Effect of Time and Sleep on the Transitive Inference Task

Effect of Time and Sleep on the Transitive Inference Task

Alexandra Morgan

A Thesis in the Field of Biology
for the Degree of Master of Liberal Arts in Extension Studies

Harvard University
March 2017

Abstract

Researchers seeking to understand the adaptive value of sleep have done so, in part, by assessing the cognitive effects of sleep on the performance of specific tasks. In this study we follow up on previous work reporting that the passage of time affects performance on the Transitive Inference (TI) task in subjects trained to below-ceiling levels on the premise pairs, and that sleep affected performance in distinctive ways. We discuss the possibility that these changes in performance are due to a shift in strategy from one in which intermediate premise pairs are coordinated as needed to respond to probes, to a strategy which relies on a gradient of preference amongst the stimuli. We argue that, in contrast to how previous work on the effect of delay and sleep on the TI task has been done, changes in performance should be measured within-subject, by testing the same subjects at multiple time points, and present evidence that this approach is valid with this task (as opposed to showing any learning effects with repeated testing.) We find, on average, no change in performance on this task over the course of 2.5 to 3 hours, with or without sleep. In contrast to previous work, we find in many subjects high levels of performance after only 20 minutes, and a pattern of performance that we expected to find only after sleep. We evaluate the use of an innovative technique to assess the preference gradient which has not previously been used with human subjects, and present evidence that the presence of a preference gradient determines how sleep affects changes in performance on this task.

Acknowledgments

Many thanks to Murray Barsky, Jon Chamberlain, Roy Cox, Herron Gomillion, Elaine Parr, Erina Sato, Anna Schapiro, and Robert Stickgold for their support, assistance, and thoughtful and useful feedback and comments.

This study was supported in part by NIH Grant MH48832. Study data were managed using REDCap electronic data capture tools hosted at Beth Israel Deaconess Medical Center, Boston. This work was conducted with support from Harvard Catalyst, The Harvard Clinical and Translational Science Center (National Center for Research Resources and the National Center for Advancing Translational Sciences, NIH 8UL1TR000170) and financial contributions from Harvard University and its affiliated academic health care centers.

Table of Contents

List of Tables

List of Figures

Chapter I

Introduction

What is the adaptive value of sleep? It is theorized that by temporarily disconnecting the brain from external stimuli and behavior, sleep allows the brain to enter an alternate information-processing mode, allowing the brain to re-organize memory in ways that result in more adaptive behavior following sleep. Sleep and other neural processes occurring during rest periods have been shown to transform memories, resulting in measurable changes in behavior over time; but on the other hand, obviously, many memory processes do function without sleep. Various tasks have been used to assess the effects of sleep on memories encoded before sleep, as reflected in behavior after sleep; but exactly how sleep alters memories has not been pinned down.

Ellenbogen *et al*. (2007) examined whether sleep had an effect on performance of a task which was not previously thought of as sleep-dependent: the transitive inference task. In transitive inference, premises are of the form A>B and B>C, where ">" can symbolize any transitive relationship; and the probe has the form A?C, a choice between items not previously seen before. This is a memory task if subjects memorize the premises at an earlier time, and respond to probes at a later time. Although this task has been extensively studied in both humans and non-human animals, before Ellenbogen et al.'s work, time had not been considered an interesting variable. It was assumed that whatever ability an animal had on the task would be apparent whenever the animal had learned the premises and was available for another testing session. Surprisingly, however,

Ellenbogen et al. reported that although their subjects responded at chance when tested almost immediately after the premise-learning session, those tested 12 or 24 hours later were significantly above chance in their ability to choose the item consistent with a transitive inference.

Most intriguingly, they reported that performance on certain pairs improved more when subjects had slept in the interval between training and test. Specifically, the sleep boost was apparent on *second degree* pairs: probes consisting of pairs of items whose relationship can only be determined by performing two inferential steps.[1] (A pair's degree of separation depends on the number of intervening items.) The effect of sleep on these pairs was replicated by Werchan and Gomez (2013).

Zeithamova et al. (2012), in trying to explain Ellenbogen's sleep-dependant results, speculate that sleep strengthens the memories for the premise pairs. However, this is not consistent with Ellenbogen's results in a couple of ways. First, if this were the case, we would expect to see improved premise pair performance at the later time points, but Ellenbogen et al. report identical premise pair performance at all time points (and Werchan and Gomez report identical premise pair performance in both the sleep and wake conditions). Secondly, why would strengthening the premise pairs benefit second degree pairs more than first degree pairs?

Models of how subjects perform transitive inference provide clues as to what might be happening. Various models have been proposed, each of which falls into one of two

[1] Given B>C, C>D, and D>E, to deduce the relationship between B and E: first infer from B>C and C>D that B>D (a first degree inference), and second, infer from B>D and D>E that B>E (a second degree inference). Thus BD is an example of a 1° pair, and BE is a 2º pair.

camps. *Coordination models* posit that when a novel pairing is presented, the intervening premise pairs are (somehow) brought into consideration and combined. Thus, such models predict that the more intervening premise pairs are required to link a novel pairing, the more computationally expensive and error-prone the novel pairing is to resolve. In contrast, *linear models*, including both *quasi-spatial models* and *preference gradient models,* claim that items linked by longer chains of premise pairs are represented in the brain as more different or widely separated in some way, and thus easier to resolve. Thus the two camps of models make opposite predictions regarding which type of inference should be performed more accurately, 1º (first degree) pairs or 2º (second degree) pairs. Curiously, in both Ellenbogen et al. (2007) and Werchan and Gomez (2013), performance on 2º pairs is lower than that on 1º pairs without sleep, consistent with a coordination model; but higher with sleep, consistent with linear models—suggesting that sleep transforms the representation of the premise pairs in the brain to support a different, more efficient computational model. Below we review various models that have been suggested, and show that a preference gradient model is most consistent with the typical pattern of performance on this task observed in both non-human animals and in humans who are not aware of an expectation that they should use transitive inference. We discuss various models of how a preference gradient has been proposed to develop over the course of training, and the difficulties these models run into. We review some of what is known about how information processing is altered in sleep, and relate this to a novel suggestion as to how an off-line process could produce results consistent with the use of a preference gradient. We go on to discuss experiments we did

to probe the possibility that an off-line process transforms transitive inference performance.

Chapter II

Background

History of the Transitive Inference task. Transitive inference is the deduction of rankings of non-adjacent members of an ordered series from the ranking of adjacent members; *i.e.*, concluding, from A>B and B>C, that A>C. Whether this inference is necessarily correct depends on the type of relationship symbolized by >. Every child knows that "rock beats scissors and scissors beats paper" does not dictate "rock beats paper". In this case the match-ups are not decided by one orderable property, and inferring transitivity would be incorrect. However, when a linear hierarchy does exist, it can be extrapolated from the outcome of a small subset of the possible match-ups. Given n items, there are n*(n-1)/2 possible pairs; but the n-1 pairs of adjacent items suffice to completely determine the entire hierarchy[2]. Thus, if enough features of the environment are intrinsically orderable, the ability to rank alternatives using transitivity has a selective advantage. By applying transitive inference, an animal can make the most use of minimal learning opportunities.

The transitive inference (TI) task can be used to probe the neural basis of this ability. In a TI task, a set of stimuli is given an arbitrary ordering. (Conventionally in the literature on this task each stimulus is referred to by a letter code corresponding with its position, with A being the highest-ranked item.) Subjects are exposed to the relative ranking within the premise pairs—the pairs of items that are adjacent in the ordering—

[2] For example, if n=10, there would be 10*9/2 = 45 possible pairs, but knowing only the (10-1) = 9 adjacent-item pairs would suffice.

and then asked to identify the higher-ranked of a novel pairing.

Originally, this was a completely verbal task, requiring that subjects understand the language in which the premises were presented, and either be able to read the language themselves, or remember information that was read aloud. In this context, transitive inference was viewed as purely a type of logical reasoning carried out by adult human beings. Piaget (cited in Bryant & Trabasso, 1971) used a verbal transitive inference task to probe the development of logical abilities in children. Children heard sentences such as "Anne is older than Betty. Betty is older than Charlie." Piaget's observation that children under the age of 7 did not reliably infer that Anne is older than Charlie led him to believe that children could not successful perform such a task until they had reached certain milestones in the development of their logical reasoning abilities.

Bryant and Trabasso (1971) revisited the question of whether younger children could perform this task, asking whether perhaps children's apparent inability was due to their inability to grasp the abstract verbal presentation of the premises and/or remember the information. To ask this, they developed the first version of the TI task in which subjects are trained to criteria on the premise pairs. In this type of task, subjects are exposed to the premise pairs many times in an initial training phase, and learn through trial and error to choose the correct item out of each pair. Testing, in which subjects choose between novel pairs of stimuli, occurs only after subjects have demonstrated sufficient mastery of the premise pairs. Thus the TI task becomes a memory task. Although immediate versions of the transitive inference task[3] do involve working

[3] Immediate versions of the transitive inference task may be verbal tasks, in which subjects read or hear the premises, or non-verbal tasks in which the premises are

memory, having to learn the premises to criteria requires that the trials leave a more durable memory trace. The question becomes, do these more durable memory traces support transitive inferences, and if so, how? Correct responses on "anchor" pairs (pairs that include either the highest- or lowest-ranked item; see Figure 2.1) may reflect a memory of nothing more than the trivial association between these items and "always winning" or "always losing" respectively. Thus, Bryant and Trabasso trained children on a 5-item series, and regarded their responses to B versus D as the critical test of transitive inference. (See Figure 2.1.)

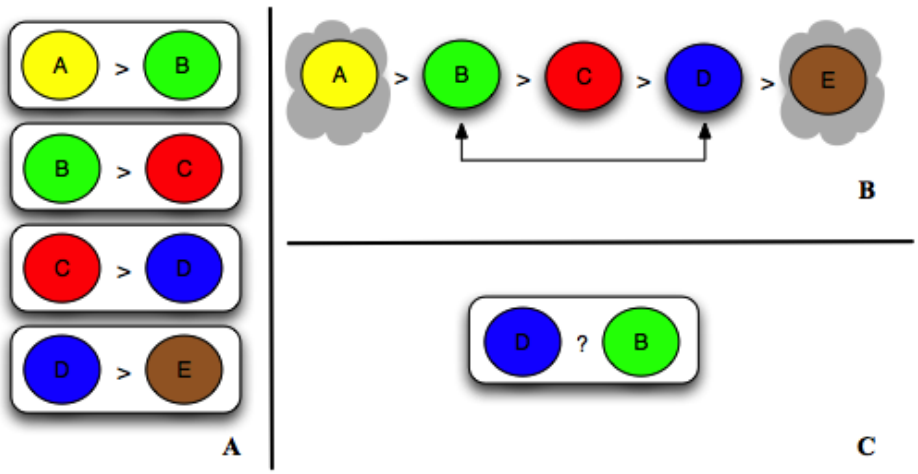

Figure 2.1: Bryant & Trabasso's (1971) TI experiment. A. Children were taught premise pairs such as these. The > symbolizes "longer than" (or another transitive relationship). Premise pairs were presented in random order, and the left/right placement of each item within each premise pair was randomized. Subjects did not see letter codes associated with stimuli. B. The implied hierarchy. Items at the end, in the grey clouds, are the "anchor" items; they always win or always lose. Arrows point out a non-anchor pairing not included in the training set. C. Example of a test trial. When asked "which one is longer?" subjects who are competent at transitive inference reply "green".

Bryant and Trabasso found that given enough training on the premises, children as young as 4 were able to correctly choose B over D. This led to the question of whether

---

available for the subjects to inspect in a graphic form, such as in Wendelken & Bunge (2010) and Mackey *et al.* (2015).

non-human animals could also perform a version of this task. After adapting this task for use with various other species, researchers have reported competent transitive inference performance in studies on apes, monkeys (Merritt & Terrace, 2011), rats (Dusek & Eichenbaum, 1997), mice, hooded crows, jays, pigeons (von Fersen, Wynne, & Delius, 1991), and fish; but not honeybees (Benard & Giurfa, 2004).

However, is a completely non-verbal version of this task truly the same task, and do subjects perform the same cognitive operations? Bryant and Trabasso had children choose between non-verbal stimuli (the ends of colored rods), but their verbal instructions and feedback encoded a relationship between what the children could see (the colors) and a hidden orderable magnitude (the length of the rods). ("Which rod is *taller*, the yellow rod or the green rod?... No, see, the yellow rod is taller.") In contrast, in non-verbal TI tasks, usually the researchers make no attempt to associate the stimuli with an orderable property. Implementations of the TI task used with animals usually depend on an *operant conditioning* procedure to imply relationships of the form "X>Y" by rewarding the selection of stimulus X when both X and Y are presented. For example, a rat learns that when presented with two small cups of sand, one scented with cumin and one scented with cocoa, digging in the cup scented with cumin (but not the one scented with cocoa) reveals half a Froot Loop (Van Elzakker, O'Reilly, & Rudy, 2003). As pointed out in Lazareva *et al.* (2004), unlike the relationship "is taller than…", the relationship "is rewarded, when presented with…" is not intrinsically transitive.[4] Thus, finding that an animal responds "correctly" on the transitive inference task, rather than

[4] Some attempts have been made in animal studies to associate stimuli with an orderable property, as in (Lazareva et al., 2004) and (Lazareva & Wasserman, 2006); evidence on whether this has an effect on behavior in non-human animals is mixed.

indicating any competence at logical operations, may merely indicate that the animal has developed preferences amongst the stimuli as a result of the operant conditioning procedure, and these preferences are transitive (Lazareva et al., 2004), a process some have dismissively referred to as "pseudoinference" (Leo & Greene, 2008).

The observation that non-verbal animals can perform a task that was previously thought of as a type of logical reasoning has inspired researchers to question how much of what appears to be logical reasoning in adult human subjects might in fact be based on phylogenetically older, non-verbal mechanisms. Higa and Staddon (1993) argued that

> …we have an obligation… to show how the complex symbolic abilities of human beings are built upon rudimentary ancestors. One way to do this is to study the conditions under which behavior that in humans would be symbolic can be brought about, by nonverbal means, in other animals.

Even when human subjects report using an explicit reasoning strategy, they may be using a nonverbal strategy that they are not aware of, as people often do not accurately report their actual decision-making processes; for example, in experiments in which subjects make a decision based on subliminally presented stimuli, subjects may confabulate reasoning processes to explain their decision *post hoc*. Thus, an elucidation of how non-human animals perform a task may shed light on the processes involved in human performance of an analogous task and vice versa. Much of the recent research on TI has sought to make the human and non-human versions of the task as analogous as possible in order to test models' predictions of behavioral results across many species. For example, Frank *et al.* (2005) replicated results of their rat study in undergraduates, and argued that the results supported their model (Frank, Rudy, & O'Reilly, 2003) as an account of the behavior of both groups; and Merritt and Terrace (2011) ran parallel series

of variants of the TI task with both rhesus monkeys and humans to test whether the same strategies are used in both species.

Models of Transitive Inference performance. Research into the Transitive Inference task has sought to address two questions. 1) How can we build a model of the algorithm in the animal's brain that supports the observed performance on this task? 2) What is the neural substrate of the implementation of this algorithm? Models or theories of Transitive Inference performance fall into three categories: coordination models, quasi-spatial models, and preference gradient models.

Coordination models posit that when the subject is faced with a novel pairing, the memories of the intervening premise pairs are activated, brought into consideration, and somehow combined. For example, when faced with B vs. D, one remembers that B beats C, and that C beats D. Logical reasoning is a coordination model: it involves explicit combination of the premises. Logical reasoning is under top-down, executive control (a function of the pre-frontal cortex in the human brain), operating on declarative knowledge. The acquisition of declarative knowledge depends on the hippocampus and other structures in the medial temporal lobe (MTL), although the retrieval of declarative memories becomes independent of the MTL over time. Reasoning can be deployed in a flexible way, thanks to the executive control. Patients with frontal lesions perform poorly on tasks such as the Wisconsin Card Sorting Task, a test requiring that the patient flexibly choose between and apply different rules for sorting cards. Top-down executive control chooses what logical rules to apply, transfers rules from one domain to another, and changes how the computation is applied depending on the question at hand. For example, a child in Bryant and Trabasso's study who had mastered knowing that the yellow rod

was *taller* than the green rod, when asked "which rod is *shorter*," would respond "green". The capacity to flexibly manipulate declarative knowledge may have evolved due to the hippocampus's original core role in encoding spatial information, which must be flexibly re-combined to find novel routes. Functional magnetic resonance imaging (fMRI) studies have shown both frontal lobe and medial temporal lobe activation in humans doing the TI task (Acuna, Eliassen, Donoghue, & Sanes, 2002; Solomon et al., 2015; Wendelken & Bunge, 2010).[5] Studies in animals have shown activation of the prefrontal cortex (pFC) (Brunamonti et al., 2016)—the cortex covering the front of the frontal lobe—and homologous structures (DeVito, Lykken, Kanter, & Eichenbaum, 2010) during TI as well.

The capacity of other species to reason in the same way as humans seems questionable, but a coordination model does not necessarily entail logical reasoning. Kumaran and McClelland (2012) propose a simple neural net, named REMERGE, which performs TI through coordination of the premise pairs. The REMERGE model has 3 layers of units: an input, or feature, layer; a conjunctive layer, whose units are meant to represent the "hippocampal" distinct, orthogonalized representations of individual facts or episodes; and a response layer. (See Figure 2.2.) During training, bi-directional connections between the feature layer and the conjunctive layer are strengthened, causing each episode and its component features to activate each other. Connections from each episode to its correct and incorrect response become more excitatory and inhibitory respectively. When the model is presented with a novel pairing—for example, when B

[5] Wendelken and Bunge (2010) did an immediate version of the TI task, and found hippocampal involvement not for performing transitive inference per se, but for encoding the relationships amongst the stimuli.

and D are activated in the input layer—all conjunctive units with either of these features (AB, BC, CD, and DE) are initially activated. The combined activation of BC and CD causes input unit C to activate, which in turn favors the activation of BC and CD over AB and DE. As BC will excite B, and CD will inhibit D, this causes greater activation of B than D in the response layer. Similarly, the model can also solve B vs. E; however, since the activation has to spread over more intermediate units, the network takes longer to settle into its stable pattern of activity, and the difference in activation between B and E in the response layer is smaller. (See Figure 2.3.)

Predictions of a coordination model oppose what is observed in most experimental data regarding the relative difficulty of pairs depending on their degree of separation. In a coordination model, a 2° pairing is more difficult to solve than a 1° pairing, because more

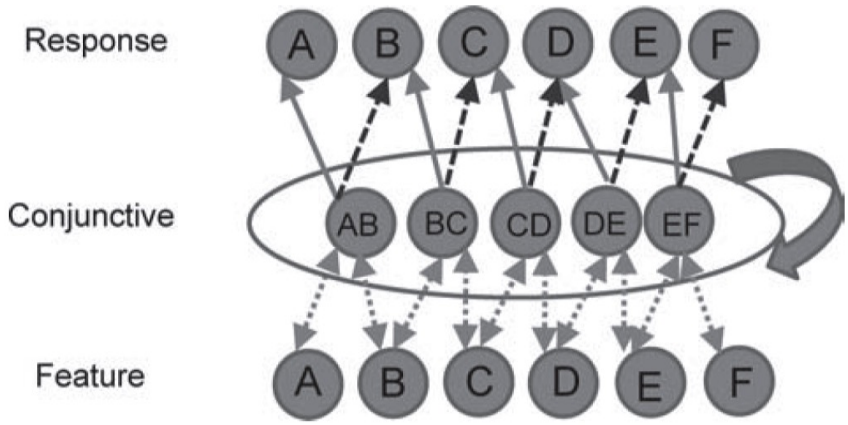

Figure 2.2: Kumuran & McClelland's REMERGE. A neural net model which displays competence at TI via re-activation and coordination of premise pairs at testing. An example of a coordination model which does not use explicit logic rules. Solid grey arrows are excitatory connections; dashed black arrows are inhibitory connections; dotted grey arrows are bi-directional excitatory connections; and the thick grey curved arrow indicates recurrent inhibitory connections within the conjunctive layer. From Kumaran & McClelland (2012).

premise pairs have to be activated and considered, and more steps are required in the processing. In contrast, consider the relative difficulty of answering these two questions: “which is further west—Nevada or Colorado?” versus “which is further west—Nevada or Kentucky?” In a spatial model, a larger disparity is more easily resolved. In most experimental data on the TI task a "symbolic distance effect" is apparent—pairs of items further apart are solved with more accuracy than pairs of items closer together.

Quasi-spatial linear models: Whereas in coordination models, the premises are not coordinated until retrieval, *quasi-spatial linear models* posit that the subject forms some kind of linear data structure in memory at the time of encoding (or at least prior to testing), integrating each stimulus into its relative position in the hierarchy. When subjects are faced with a novel pairing, they then simply look up the relative positions of the stimuli in the hierarchy. Some (Gazes, Lazareva, Bergene, & Hampton, 2014; Jacobs, 2006; Merritt & Terrace, 2011) speculate that the hippocampus's innate spatial abilities

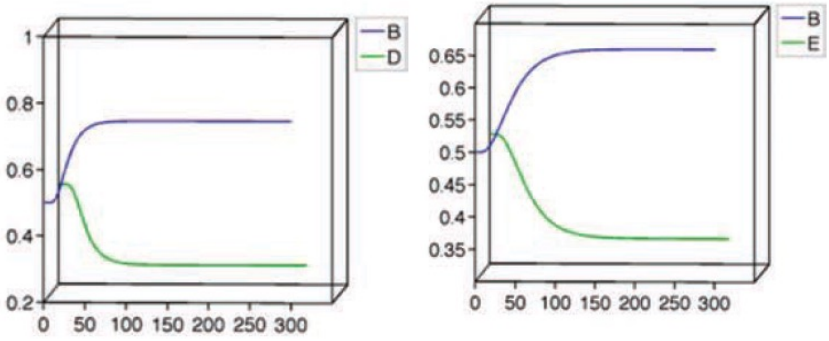

Figure 2.3: Response layer activations in the REMERGE model, when the model is presented with B vs. D (left) or B vs. E (right). Notice the differing axes: the model activates item B more strongly when paired with D (a 1º separation) than when paired with E (a 2º separation). From Kumaran & McClelland (2012).

have been co-opted to manipulate data in a different domain in an analogous way. This implies, however, that mapping arbitrary quantities to a spatial arrangement is innate. If this is true, it is surprising that this practice was not common in mathematics until the 17th century. When John Wallis (1616-1703) extended the number line into the negative direction, and when Rene Descartes (1595-1650) placed two number lines perpendicular to each other to represent pairs of numbers, these innovations were not immediately seen as natural, but as revolutionary and even controversial. This suggests that perhaps the mapping of arbitrary quantities to lines is not innate, nor intrinsically obvious and automatic, but only seems natural to highly educated human beings who have been presented with this way of representing quantities since grade school. Human minds readily take to the idea of making an analogy between relative quantities and a linear arrangement in space; but there are many examples of equivalences so embedded in our thinking (such as "argument" is "battle", "communication" is "shipment of goods", etc.) that we do not even recognize when we are using metaphor (Lakoff & Johnson, 1980). This is probably a uniquely human particularity, a by-product of our use of language. Of course, it is highly educated humans, well versed in the conventions of modern, post-Descartes mathematics, who are positing formation of mental number lines in the brains of animals doing the TI task.

A quasi-spatial arrangement is not the only data structure that could produce transitive performance with symbolic distance effects. A quasi-spatial model is analogous to, in computer science terms, an array: in a digital computer, pointers to each stimulus would be laid out in adjacent memory locations, in the same order as in the actual hierarchy. This is not the only way the ordering could be stored. If a hash table is used,

pointers to the stimuli are laid out in an arbitrary order; but associated with each of these is a number. These numbers encode the order of the hierarchy, if their rank orders correspond to the positions of the associated stimuli. (Of course, the architecture of a digital computer is utterly different from that of a nervous system, so this is merely a high-level analogy, to demonstrate the algorithmic feasibility of each arrangement.)

Preference gradient models posit that each stimulus becomes associated with a "value" or "excitatory strength", and that the gradient of these values down the hierarchy, from best to worst, is an artifact of the operant conditioning procedure used to train the subject. Conditioning is a type of memory that is more slowly acquired than declarative memory, and influences behavior in a less flexible way. The neural substrate of reward-conditioned behavior is domaminergic corticostriatal circuitry involving the basal ganglia. Dopaminergic neurons in the striatum fire when a reward (either intrinsic, such as juice, or extrinsic, such as money) exceeds expectations, thus updating the reward contingencies encoded in the corticostriatal loops. As this mechanism associates a stimulus with a reward, approach behavior towards the stimulus increases. The hippocampus is involved in the reward circuitry as well; learning by the hippocampus may be dopaminergically modulated, and the hippocampus mediates the association between reward and context. (For a general review of reward-related learning as it impacts decision-making, see Delgado & Dickerson (2012).)

Experiencing rewards during the TI training procedure undoubtedly engages an animal's reward-processing circuitry when the reward is as salient as half a Froot Loop must be to a rat. However, competent performance on the TI task requires a more complex calculation than simply "approach the more-rewarded stimulus." In most

versions of the task, during training, the subject experiences exactly the same number of trials in which B yields a reward (BC trials) and trials in which it doesn't (AB trials); and likewise for the number of DE vs. CD trials. Thus the expected values of stimuli B and D are equal, both having a 50% chance of reward (assuming a 5-item or longer series).

Various models have been proposed to try to explain the observed behavior on the TI task as a function of reward history. One of the earliest, and most prominent, models to try to explain TI performance based on reinforcement history is the Value Transfer Theory (von Fersen et al., 1991). This theory posits that the effective value of each stimulus is the sum of the reward value of the stimulus and some function of the effective value of the other stimuli with which it was presented.[6] Since B is sometimes presented with A, its effective value is a function of the reward value of A; and since that is higher than the reward value of any other stimulus, the effective value of B is inflated compared to other stimuli down the line; B, in turn, inflates the value of C, and so on. Thus a gradient of value develops from the transfer of value down the hierarchy. The differences in values assigned to stimuli by this theory was observed to match the accuracy of pigeons, trained on a 5-item series, in choosing the higher-ranked item in a novel pairing (von Fersen et al., 1991); and also predicts that the value curve is steeper at the ends, and thus the observed *serial position effect* (SPE). The SPE is the typical U-shaped curve in performance on the premise pairs, with better performance on premise pairs at the ends

[6] Based on consideration of asymmetries in premise pair performance, von Fersen et al. concluded that value transfers only from more-valued stimuli to less-valued stimuli, and thus only in one direction along the hierarchy. However, in species other than pigeons, the asymmetries amongst premise pairs are sometimes different. Must the mechanisms of reward-based learning be different in different species, such that value transfer is observed in all species, but in different directions? This seems to be a weakness of the Value Transfer Theory.

than those in the middle (Figure 2.4).

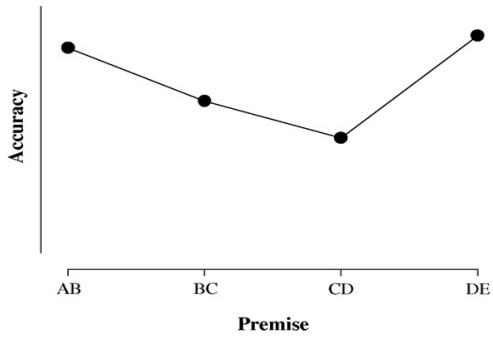

Figure 2.4: The serial position effect, the typical U-shaped curve in premise pair performance as a function of hierarchy position. From Vasconcelos (2008).

Vasconcelos (2008) has pointed out, however, that the Value Transfer Theory does not model a process. It predicts, mathematically, the relationship between the valuations of the different stimuli at the end of training, but does not predict how these valuations update on a trial-by-trial basis. Since the pigeon experiences training trial by trial, learning over the course of each trial, rather than having summary statistics uploaded into its brain all at once, this is a huge hole in the theory. Furthermore, the form of the theory that matches the pigeon data—in which value transfers just from the higher-ranked item to the lower-ranked item in any premise pair—predicts that we would see extremely poor performance on the CE novel pairing in a six-item series. Data from pigeons (Daniels, Laude, & Zentall, 2014), rats (Van Elzakker et al., 2003), and unaware humans (Frank et al., 2005) trained on six-item series have shown, if anything, better performance on CE than on BD.

More sophisticated algorithmic and neural models of TI performance in animals have generally included both *elemental* and *conjunctive* encodings of the stimuli. An elemental encoding updates the value of each stimulus individually, whereas conjunctive encoding updates the value of each stimulus in the context of both stimuli appearing together; *i.e.*, after presentation of a B vs. C trial, how much does the brain update its representation of B itself, versus its representation of how to respond to B given the

combination of B and C together? Neither type of encoding alone can explain animals' observed performance on this task. An animal attending only to the elemental value of the stimuli would fail to learn the premises. The selection of B is rewarded and not rewarded equally frequently, and the same could be said for stimulus C. Thus, disregarding context leaves the animal with no basis for choosing B over C. On the other hand, an animal who only learned conjunctive representations of the stimuli would not be expected to respond transitively to novel pairings: a BD trial is a novel context, so without any representation of the stimuli independent of context, there is no basis for choosing between B and D.

Various models differ in how these two types of representation work together to support observed performance. In the *eta-kappa model* developed by Delius and Siemann (1998), each trial updates both V(X) (the elemental value of stimulus X) and V(X | XY) (the value of X in the context XY). Performance on premise pair trials is explained by some combination of the elemental and conjunctive value of the stimuli. Some species, such as humans, rely more on conjunctive encoding for premise pairs, and are thus able to learn the premise pairs of a circular structure (such as rock-paper-scissors), whereas others, such as pigeons, seem to be less able to rely on conjunctive encodings, and have great difficulty learning a circular set of premise pairs. In either case, transitive inference develops due to the relative elemental value of the stimuli.

The hippocampus, particularly the dentate gyrus, is exceptionally well suited to the formation of conjunctive representations due to its sparse representations of input patterns, resulting in distinct, separate encodings of patterns with overlapping features. This results in the preservation of idiosyncratic features, at the expense of finding

generalities. In contrast, non-hippocampal cortex finds regularities in the environment by overlapping its representations of patterns with common features; *i.e.*, every time cortex encounters an input pattern containing the feature "isa-bird", synaptic weights related to a single "isa-bird" representation are updated, whereas the hippocampus will represent "sparrow isa-bird" and "penguin isa-bird" in entirely disjoint patterns of activation. According to Complementary Learning Systems (CLS) theory (McClelland, McNaughton, & O'Reilly, 1995), this feature of the hippocampus is important for circumventing a key drawback of the cortex's unified representation. The cortex, on its own, cannot quickly find an accommodation of an anomalous pattern; *i.e.*, quickly updating the network to learn that a penguin is a bird and that it swims and does not fly wrecks the network's prior understanding that birds fly. The only way the cortex can accommodate the "penguin" and "sparrow" patterns, given a unified "isa-bird" unit, is to gradually explore representation-space to find a distinct place for the penguin concept. This requires gradual learning through many presentations of the "penguin" pattern, interleaved with continued repetitions of the other "bird" patterns. However, the environment is not so kind as to provide just the right training schedule: animals must often quickly learn an anomalous fact given one learning opportunity. The CLS theory's solution to this is to have the hippocampus as a temporary holding pen for idiosyncratic memories, which are then gradually replayed back to the cortex over time so that the cortex can find a generalized representation that accommodates all exemplars. This special role for the hippocampus may offer clues to why the hippocampus is required for normal TI performance. Analogous to the hippocampus's role in keeping seemingly contradictory facts such as "most birds fly" and "penguins do not fly" distinct until the

cortex finds a way to reconcile their representations, perhaps it also keeps “D loses when paired with C” and “D wins when paired with E” distinct until these facts can be incorporated into an overall representation.

A common feature of almost all prior work on the TI task is that it has not considered the possibility that the memory traces may change over time, after the learning experiences, producing improvements in performance. In almost all studies involving human subjects, testing on the non-trained pairs immediately follows training. Failure to respond correctly on the non-adjacent pairings in an immediate test is taken as indicating that whatever learning the subjects achieved was not sufficient to support successful performance on the transitive portion of the task, period. In animal studies, because of the more extensive training required for less encephalized animals to reach criterion on the premise pairs, training commonly takes place over the course of many days; however, time elapsed between training and testing has not generally been considered an interesting independent variable in these studies.

Sleep and memory: This is in spite of the growing realization that off-line time between experience and performance, and brain state during that time, are key factors in many types of memory. Although in common usage, the word “memory” refers to the conscious experience of either mentally re-living prior events or recalling factual material, memory researchers define memory much more broadly as “the effects of experience that are manifest at a later time.” (Gallagher, 1990) Thus it encompasses not just the human experiences of mental time travel—episodic memory—and the ability to regurgitate facts—semantic memory—but also procedural memory, Pavlovian conditioning, priming effects, conditioned taste aversion, etc., all the way down to gill

sensitization in sea slugs, all of which may be encoded in different ways by the nervous system.

Given the diverse roles different types of memory have in enhancing an animal's chances of survival, various types of memories have different requirements for how they should be transformed over time. We tend to think of the optimal effect that time should have on memory as "preserve it as exactly as possible as long as possible." After all, our most conscious experience of grappling with our own memory is the experience of trying to recall the contents of a textbook at exam time as well as when we were staring at the pages. However, this is as naïve and over-simplified as thinking of "memory" as just "conscious recall of events or facts". Facts and events do need to be recalled veridically sometimes, but they may be more useful when compared and combined to discover generalizations; distinct instances of events of the same type need to be combined to discover schemas, without losing the distinct details of each event; detail may be lost in favor of gist; motor skills improve through becoming more automatic. Emotional tone may be useful to tag memories for enhanced encoding, but once the life lessons of a trauma have been extracted, it is best that the emotional color fades. As many have pointed out (Eichenbaum & Fortin, 2009), the adaptive significance of memory is not to capture the past *per se*, but to optimize future behavior.

Given the diverse requirements for how different types of memory should evolve, and the diverse neural mechanisms underlying their evolution, it is not surprising that findings on the effects of time and brain state on memory have been diverse. The earliest findings on the effect of sleep on memory by Jenkins & Dallenbach (1924) was on the verbal declarative memory of strings of nonsense syllables. They found that the decay in

memory over time occurred at a slower rate during sleep. More interesting transformations of memory over time are possible when the information is meaningful. Later work has found that emotional or meaningful information does not just decay, but may be transformed to extract the gist (McKeon, Pace-Schott, & Spencer, 2012) or most salient features (Payne, Stickgold, Swanberg, & Kensinger, 2008). Procedural memory—both motor (Walker, Brakefield, Morgan, Hobson, & Stickgold, 2002) and perceptual memory (Stickgold, Whidbee, Schirmer, Patel, & Hobson, 2000)—may actually improve over off-line time, especially with sleep.

Over the past several years, dozens of labs have sought to define exactly which memory processes are benefited by sleep (Conte & Ficca, 2013). Whether or not, and how, a memory is transformed by sleep may depend not only on its type (*i.e.* declarative vs. procedural), but also on subtle differences in a subject's learning experience. These include whether or not subjects are explicitly aware of what they are learning; for example, sleep is required to enhance performance on the Serial Reaction Time Task only when subjects are aware of the repeating pattern of button presses (Robertson, Pascual-Leone, & Press, 2004). Another factor is whether subjects have extensive experience in a domain: while most human subjects show a boost only after sleep in performance on the motor sequence task, experienced musicians show improvements after a 12-hour delay regardless of whether they slept (Tucker, Nguyen, & Stickgold, 2016). Also relevant is the difficulty of the task: for example, sleep enhances performance on a repeated presentation of remote associates puzzles, but only for the more difficult items (Sio, Monaghan, & Ormerod, 2012).

To what extent the survival value of sleep has to do with its role in off-line information processing is unknown. Clearly sleep must have survival value; it is evolutionarily conserved, being found in some form in all animals with brains; furthermore, robust mechanisms have evolved to homeostatically regulate both the amount of time animals are asleep and the spectral power of brain activity in frequency bands characteristic of deep sleep. The arguments that sleep serves the trivial purpose of keeping animals immobilized for energy conservation or to avoid predation fail to take into account the fact that energy intake or expenditure has very little effect on sleep need, and that animals for whom avoiding predation is difficult during sleep still sleep. Even marine mammals who must swim continuously to avoid drowning have evolved a way to sleep: the two hemispheres of their brains take turns napping (Mukhametov, Supin, & Polyakova, 1977).

Exciting new research has shown that sleep may serve the important physiological purpose of clearing out misfolded proteins and other cellular debris in the brain (Xie et al., 2013). However, sleep has several distinctive electrophysiological features whose role in such a process seems hard to imagine, but which, on the other hand, do seem to have something to do with information processing. Spindles, bursts of 12 to 15 Hz EEG activity lasting 1 to 1.5 seconds, a defining feature of stage 2 and deeper non-REM sleep, have been linked to both synaptic plasticity and coordinated hippocampal replay of learning experiences (Marshall & Born, 2007). Localized spindle density has been found to be correlated with several types of learning, such as the motor sequence task (Nishida & Walker, 2007) and visual perceptual learning (Bang, Khalilzadeh, Hamalainen, Watanabe, & Sasaki, 2014).

Replay of spatial paths, the firing of sequences of cells coding for spatial locations in the order in which they were visited while traversing a maze, has been observed in the rat hippocampus during sharp wave ripples, which occur during slow wave sleep (SWS) as well as in awake rest states. Gupta *et al.* (2010) found, in a more complex environment, replay of sequences of spatial locations corresponding to paths that the rat had never taken. Some of these sequences were backwards replay of routes the rat had taken, while others were novel routes formed by re-combining portions of previously experienced paths. They theorize that by, in effect, simulating alternate routes, this re-combination allows the rat to learn a complete map of the environment. Here we have electrophysiological evidence that off-line rest periods, specifically those which contain sharp wave ripples, allow the brain to explore combinations of elements not actually experienced before. Gupta *et al.* only monitored this hippocampal activity during awake rest states, but given the finding of sleep-dependent improvements in navigating a maze (Nguyen, Tucker, Stickgold, & Wamsley, 2013) it seems likely that this re-combination process is even more productive in sharp wave ripples during sleep.

Conjoining different memories allows an animal to discover new routes in a spatial paradigm, and also allows an animal to come to correct conclusions in an inference paradigm. Associative inference is a similar task to transitive inference, requiring the subject to conclude from "A goes with B" and "B goes with C" that "A goes with C". There is evidence for off-line processing supporting an associate inference task: Schlichting and Preston (2015) report evidence of increased communication between the hippocampus and medial prefrontal cortex following the learning of associations that overlap with previously learned associations, and that this increased communication

correlates with subsequently improved associative inference, suggesting that a prefrontal-hippocampal dialog during rest periods plays a role in integrating memories.

The most anomalous feature of sleep in terms of neural activity, found in homeotherms (mammals and birds), is rapid eye movement (REM) sleep. This stage of sleep is also known as "paradoxical sleep", because the brain is as active in REM as during wake, and the EEG is similar. Although normally REM can only be entered from other stages of sleep, to some extent REM is controlled independently of other stages of sleep: REM propensity is highest at a different circadian phase than is the pressure for deep sleep, and the amount of REM sleep within sleep is under some homeostatic control, as suppression of REM will result in a rebound in REM intensity later. This suggests that REM has functions that are distinct from those served by other stages of sleep.

Boyce *et al.* (2016) have experimentally demonstrated a role for REM sleep in hippocampally-dependent memory. In mice, hippocampal theta rhythms during REM sleep were blocked through optogenetic silencing of GABAergic neurons in the medial septum for 4 hours after learning sessions. Unlike control mice, the REM theta-blocked mice showed no preference for exploring an object whose location had changed, compared to an unmoved object; and showed reduced freezing behavior when placed in an environment in which they had received a shock (but normal freezing behavior in response to a tone that had been paired with a shock). Thus, some process in the mouse's hippocampus, crucial for later expression of hippocampally dependent memories, occurs during REM sleep and is driven by theta oscillations.

A tantalizing hint that REM may be the brain's chance to process information in some completely different way than in wake is what we often recall when awoken from

REM: dreams which are more vivid, bizarre, and narrative than from other stages of sleep. Much of the bizarreness of REM dreams comes from the incongruent re-combination of narrative elements (people, places, things, and motivations): Santa Claus in the bathroom, neon lights over the high school lockers, and Dr. Stickgold moving to Peru to direct a lab that runs a radio station are typical examples of dream incongruities. Thus, phenomenologically, there is evidence for memory recombination during REM.

Intriguing experimental evidence that the human brain processes information in REM in a distinctive way includes the finding that when subjects are awakened from REM sleep, and tested immediately after awakening while the brain's neuromodulatory state is still similar to REM, the priming effect of weakly associated words is paradoxically stronger than those of strongly associated words (Stickgold, Scott, Rittenhouse, & Hobson, 1999). Creativity has been found to be enhanced by REM (Cai, Mednick, Harrison, Kanady, & Mednick, 2009). Could it be that one of the purposes of REM sleep, as suggested by Walker and Stickgold, is to support the "integrative stage of memory processing" (Walker & Stickgold, 2010), allowing the brain to develop responses that require the coordinated consideration of multiple experiences? Consistent with this theory, improvement on a *probabilistic* categorization task—which requires observing many trials to absorb the probabilistic associations between stimuli and outcome—has been found to correlate with REM sleep (Barsky, Tucker, & Stickgold, 2015). This predicts that REM would benefit a task requiring more integrative than veridical recall of the training, such as the Transitive Inference task.

A proposed off-line process supporting TI performance. I propose that there is an off-line process that runs during rest or sleep which processes memories of the form "A

was rewarded over B". In this hypothesized offline process, the brain reactivates memories of stimuli, but not necessarily paired up in the pairs encountered during training; rather (given the brain's tendency to recombine memories during sleep) random selections of stimuli are recalled. A coordination process determines which stimulus is more favored out of the imagined pairing. Thus, in rumination, the animal may encounter any of the possible pairings: AB, AC, AD, etc... BC, BD, BE, etc... DE, DF, and EF. Out of all possible pairings, in a 6-item TI task, A will win 100% of the time, B 80%, C 60%, and so on, based on chaining together the intermediate premise pairs. Thus, if this off-line simulation is run enough, the animal "experiences", in its ruminations, a gradation of expected value from highest to lowest across the hierarchy. As long as the expected value always decreases across the hierarchy, transitive performance is predicted (because E(B) > E(D)) and with a symbolic distance effect (because E(B)-E(E) > E(B)-E(D)).

Prior work on the effect of sleep on the TI task: Although they did not have any proposal such as the one outlined above for why this would occur, Ellenbogen *et al.* (2007) hypothesized that the memory-processing role of sleep would enhance or alter performance on the Transitive Inference task. Their subjects received minimal training on the premise pairs of a six-item sequence. They were then tested either after 20 minutes, 12 hours awake, 12 hours including a night of sleep, or 24 hours. After 20 minutes, subjects were entirely incapable of transitive performance, responding on average at chance level on all inference pairs. After 12 or 24 hours, however, the ability to make transitive inferences had developed; subjects responded significantly better than chance on all inference pairs. Intriguingly, the ability to answer correctly on pairs separated by two intervening items (second degree (2°) pair B-E) improved more in subjects who slept

during the interval between training and test—more than for 1º pairs and more than subjects who were awake between training and test—suggesting a sleep-dependent shift away from having to coordinate premise pairs at retrieval, towards the development of a preference gradient. (See Figure 2.5.) This effect of sleep on the 2º pairs was replicated, using a nearly identical version of the task, by Werchan and Gomez (2013).

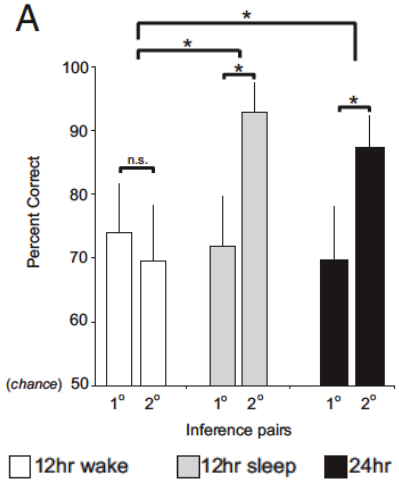

Figure 2.5. Post-sleep TI performance. (Ellenbogen, Hu, Payne, Titone, & Walker, 2007)

Neither Ellenbogen et al. nor Werchan and Gomez measured the sleep parameters of their subjects using polysomnography. Other tasks have shown correlation between improvement and time in specific sleep stages (Stickgold et al., 2000) or electrophysiological characteristics of sleep (Nishida & Walker, 2007); these findings guide attempts to understand what the brain does with memories during sleep. When sleep stage correlation studies can be done using midday naps, more subjects can be run for the same amount of resources, increasing the power of these studies; and a nap avoids the question of circadian confounds. Naps tend to be very diverse in their sleep stage composition. Greater variance in the amount of each sleep stage experienced by each subject enhances a study's ability to detect sleep stage/performance correlations. Nonetheless, some studies report a sleep vs. wake effect on performance, without identifying any significant specific sleep stage correlation (Nguyen et al., 2013). For this reason, in the study described below, we compared the performance of subjects who took a PSG-monitored nap with

subjects who were awake for a similar length of time. Ellenbogen *et al.* reported a change in TI performance between 20 minutes and 12 hours even without sleep, so to pin down the early time-course this study looked at evolution of TI performance at an intermediate time point in the wake subjects as well.

In animal experiments, animals are trained over the course of many days. Thus there is plenty of down-time over the course of training for the off-line process to occur and a gradient of expected value to develop, which could explain the fact that the symbolic distance effect is seen from the start of testing. The fact that single-session learning of the premise pairs is possible with human subjects motivates doing this study with humans rather than with other animals, even though the memory-recombination function of sleep is likely evolutionarily conserved.

Human subjects present other challenges. Often all scores are at ceiling across the board. In many of the previous studies, human subjects have usually either a) explicitly known or surmised that there is a linear hierarchy, either from the instructions, or the observation that the premises are consistent with a linear hierarchy; or b) are over-trained (trained until they are at ceiling on the premises). In cases where they are aware that there is a linear hierarchy, explicit logical reasoning can explain the subjects' performance. Logical reasoning is one of the most reliable coordination models. Frank *et al.* (2005) found that of their 65 subjects who learned the premise pairs, 8 subjects were explicitly aware of the hierarchy and thus that transitivity would apply; and these 8 subjects were at ceiling in all types of probes. The near-ceiling performance produced when human subjects use a logical reasoning strategy is likely to obscure any improvements due to off-

line development of a preference gradient. Thus, any study using human subjects has to try to assess awareness.

Explicit awareness of hierarchy: The assumptions that humans who are aware of the hierarchical relationship between the stimuli rely on logical, explicit reasoning, whereas those who are not aware rely on strategies similar to those used by rats and pigeons, has led to contention as to how to measure awareness, and exactly what we mean by the word "awareness". Using more or less stringent criteria for judging "awareness", some researchers have reported a correlation between subjects' TI performance and their self-reported awareness of a hierarchy amongst the stimuli (Frank et al., 2005; Martin & Alsop, 2004; Smith & Squire, 2005), whereas others have reported no correlation (Greene, Gross, Elsinger, & Rao, 2006; Greene, Spellman, Dusek, Eichenbaum, & Levy, 2001) Even if human subjects report using logical reasoning based on their awareness of the task demands, this is not necessarily reliable; humans may use the same algorithm to provide an answer as do the rats, with their verbal abilities merely supplying a *post hoc* justification for their answers. Greene *et al.* (2001) found evidence for the development of task awareness after task competence developed. Regardless, it is certainly the case that factors that increase the likeliness of subjects reporting "awareness" also increase their ability to do the task (Kumaran & Ludwig, 2013; Lazareva & Wasserman, 2010). Various factors in how the training is implemented influence awareness, such as whether training trials are organized in blocked or randomized order (Greene, 2007). The fact that there is usually[7] no symbolic distance

[7] Although there are exceptions. In Acuna, Sanes, & Donoghue (2002), symbolic distance effects were observed even though subjects were explicitly aware of the hierarchy.

effect when subjects report a high degree of awareness of the structure of the hierarchy is consistent with the possibility that they may be relying on explicit logical reasoning, as this is a coordination model.

It is unlikely that many of the subjects in either of the studies reporting sleep effects on TI (Ellenbogen et al., 2007; Werchan & Gomez, 2013) were using logical reasoning to perform the task. Subjects were not told that there was a hierarchy nor told of any transitive relationship amongst the stimuli. Other parameters of the training protocol were those that have been shown to minimize task awareness: the stimuli were abstract visual patterns, which are difficult to encode verbally (see Figure 3.2); training trials were presented in intermixed order; and subjects were trained to a lower criterion of mastery than in most previous studies. TI scores in these studies were still not at ceiling after sleep, unlike subjects who reported the use of verbal reasoning strategies in previous TI studies. Furthermore, Werchan and Gomez excluded those subjects (10% of subjects recruited) who reported awareness of the hierarchy on a post-testing awareness questionnaire. Presumably subjects who report using explicit logical reasoning would develop TI ability sooner; however, this has not been tested because Werchan and Gomez did not test TI abilities within less than 12 hours in any subjects. Ellenbogen *et al.* did not administer awareness questionnaires, but did collect confidence ratings. They argue that the increased TI scores over time are not due to increased awareness, based on the fact that confidence ratings did not increase. However, confidence is a dubious proxy for "awareness". One can be confident in responses that were arrived at by processes other than explicit application of transitive inference; or, one could explicitly apply transitive

---

However the hierarchy was unusually long (11 items); because of this, use of explicit reasoning strategies may have presented too large a burden on working memory.

inference, but not be sure of the correctness of doing so. Thus, neither of these studies conclusively nailed down the effect of awareness on the time course of TI performance, or vice versa. We hypothesized that subjects who displayed great "awareness" on a questionnaire would be at ceiling on transitive inference very soon after training, but that others who did not could develop transitive inference over time.

Although the performance of subjects in both Ellenbogen et al.'s study and Werchan and Gomez's study was not similar to those in studies in which subjects were explicitly aware of the transitive relationships, neither was it similar to results previously obtained with animals or with non-aware human subjects—at least not until after subjects had slept. In contrast to previous studies on TI, Ellenbogen *et al.*, and Werchan and Gomez, saw symbolic distance effects only when subjects had slept between training and testing. Their subjects who spent 12 hours awake were worse at the 2° pair than on either 1° pair. Furthermore, most previous studies show a series position effect on the premise pairs. Ellenbogen's subjects do not show this U-shaped performance curve; in fact, their immediate performance on AB is worse than other premises. If subjects remember that A is always rewarded, or that F is never rewarded, they should be able to trivially choose A without inference; and indeed, in prior studies, subjects have performed extremely well on the pair composed of the items at the ends. After 12 hours, Ellenbogen's subjects had higher scores on AF than on either AB or EF, as would be expected; but their scores on AF were surprisingly low after only 20 minutes.

These results are not consistent with the immediate use of elemental value of each stimulus, but are consistent with a delayed, post-sleep development of a gradient of value across the hierarchy. Based on the Ellenbogen results, we hypothesized that human

subjects who were trained only enough to have a tenuous grasp of each premise pair, not to ceiling, would initially mostly rely on the conjunctive value of the stimuli; *i.e.*, although they would be able to respond correctly on the premise pairs, scores on any novel pairing would be low. After a delay, either awake or asleep, reliance on elemental value of stimuli would increase, reflected in an increased score on the end-item pair, but novel pairings involving only non-anchor items would require coordination of premises at retrieval, and thus not show symbolic distance effects. The off-line process would occur more slowly than in experiments involving more intensive training; but after sleep, the off-line process would have produced enough of a gradient of value amongst the hierarchy members to support transitive performance, and symbolic distance effects would be apparent, reflected in better accuracy on BE compared to the 1º pairs.

Thus, although performance on 1° pairs is similar after 12 hours either awake or asleep, the mechanism by which these judgments are made may be markedly different. If E(X) were measured directly—for example, using forced choice between hierarchy members and a novel item—the rate at which the trained item is selected should vary across the hierarchy when there is a preference gradient. However, existing work has rarely tried to test the associative strength of individual members of the trained hierarchy. The work described below seeks to establish whether a *novel item test*—a set of test trials pairing each of the stimuli in the hierarchy with a previously not encountered, novel stimulus—is correlated with the symbolic distance effect.

Ideally, it would be most informative to probe both the transitive inference abilities and elemental stimuli values in the same subjects over time, unless a prior exposure to the TI probe trials, or the novel item test, would alter or disrupt the normal course of

subjects' memories over time and sleep. It is unfortunate that in both Ellenbogen et al.'s study and in Werchan and Gomez's study, the subjects who were tested on inference pairs 12 hours after training were not also tested after only 20 minutes. We presume that they had improved over the 12 hours, because a separate set of subjects, recruited in the same way from the same population as Ellenbogen's 12-hour subjects, and trained on the premise pairs using the same computerized task, were at chance when tested after only 20 minutes. But we cannot exclude the possibility that the 20-minute group overrepresented subjects who could not do transitive inference, and the 12-hour groups overrepresented subjects who could, for unknown reasons. Without having tested the same subjects at 20 minutes, improvement between 20 minutes and 12 hours has not been proven. But repeated measures is not possible in tasks where there is a learning effect, or in tasks where subjects' exposure to the test alters their expectations regarding the future usefulness of what they have learned. As prior work had not tried to do repeated measures on this task in humans, the work described below attempts to settle this point in a first experiment, before using a repeated-measures design in a second, nap-based experiment.

Chapter III

Materials and Methods

Experiment 1

Experiment 1 was designed to test whether transitive inference ability develops over a short (2.5 to 3 hour) period awake, and whether repeated measures are valid with the TI task. Subjects trained to criteria on the premise pairs of a six-item hierarchy. Some subjects were tested on transitive inference (TI) pairs at both 20 minutes and 3 hours after training. Others were tested only at 3 hours. All subjects remained awake between training and the 3 hour test. See Figure 3.1 for the protocol timeline.

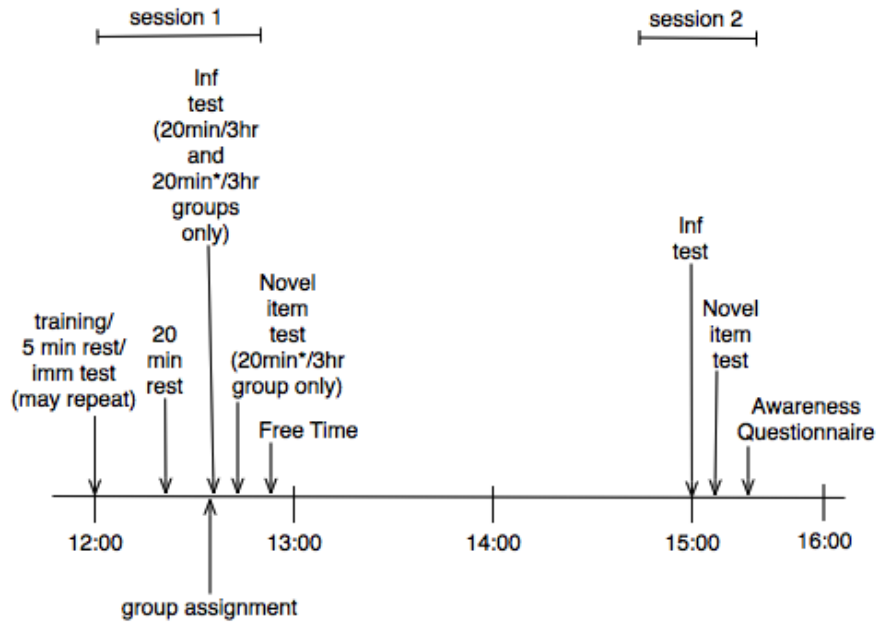

Figure 3.1: Timeline of protocol for Experiment 1.

Participants: A total of 37 subjects (27 female) between the ages of 18 and 30 (mean 21.4, standard deviation 3.0) were recruited by posting recruitment ads to the on-line job boards at local universities, by inviting subjects who had previously participated

in other studies at the Center for Sleep and Cognition, and by posting tear-off flyers near the Beth Israel campus. Potential subjects were screened using an on-line surevy in REDcap, a secure database for health-related research data developed at Vanderbuilt University (Harris et al., 2009). All subjects reported having a typical bedtime between 10 p.m. and 2 a.m., and a typical time in bed between 6 and 10 hours. By self-report, none had ever been diagnosed with a psychiatric, neurological, or sleep-related disorder, nor were they currently on any psychoactive medication. Subjects were required to keep a regular sleep schedule for 3 nights prior to the study visit day, abstain from alcohol and recreational drugs the day before the study, and cease consuming caffeine by 10 am of the study visit day. Subjects gave written informed consent approved by the Beth Israel Deaconess Medical Center Institutional Review Board and were paid in cash for their participation.

Task: The Transitive Inference task as described in (Ellenbogen et al., 2007) was re-implemented in Java and run on PCs running Windows 7 at the Center for Sleep and Cognition at Beth Israel Deaconess Medical Center.

Eight colorful ovals, shown in figure 3.2, were used as stimuli. Six of these were randomly selected and assigned positions in a hierarchy from A (most favored) to F (least favored); the remaining two were assigned to be X and Y in the novel-item test. As there are over 40,000 ways to order 8 items, the assignment of ovals to positions differed for each subject. Before the start of training, subjects saw the following instructions on-screen:

> In this task, you will see two abstract patterns at a time. Your job is to decide which of the two patterns is covering a smiley face.
>
> If you think that the pattern on the LEFT covers a smiley face,

press the key labeled 'left'. If you think that the pattern on the RIGHT covers a smiley face, press the key labeled 'right'.

At first, the task will seem very difficult, but over trials it will become easier.

Immediately afterwards, before each round of training, subjects saw these instructions:

Please keep your index fingers over the 'left' and 'right' keys during the task and respond as quickly and accurately as you can.

After you make your response, the smiley face will appear if you are correct. If you are incorrect, the smiley face will not appear.

GOOD LUCK!

For the first 18 subjects, only one round of training was allowed (see below); and for these subjects, the start-of-task instructions and the pre-training instructions were concatenated onto one screen.

All training was randomized-order training. This training was organized into blocks of 10 trials, consisting of two presentations of each of the 5 premise pairs, in a pseudorandom order in each block. Consecutive trials during randomized-order training never featured any of the

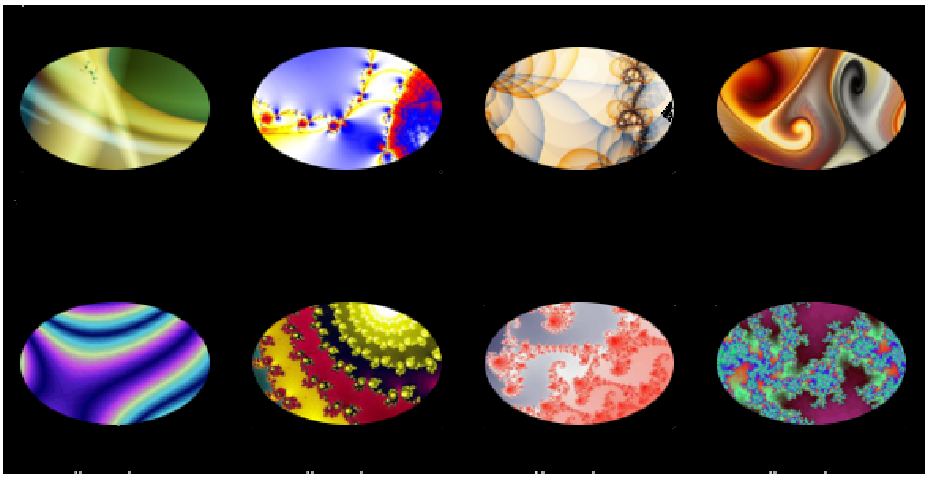

Figure 3.2: Stimuli used in the Transitive Inference task

same stimuli. Each block had a different pseudorandom order, but these orders were the same for every subject. Within each block, each stimulus appeared an equal number of times on the left and right sides of the screen.

On each trial, once the two stimuli appeared on screen, subjects were allowed unlimited time to make their selection. They selected the stimulus either to the right or to the left by pressing corresponding keys on a keyboard. On each training trial, after the subject guessed which item is correct, the selected item moved to the top of the screen to reveal either a smiley face indicating a correct choice, or nothing, indicating an incorrect choice. The first round of training consisted of at least 4 blocks (*i.e.*, 40 trials), and subsequent rounds consisted of at least 2 blocks (*i.e.*, 20 trials). Each round of training continued until the subject responded correctly at least once to each premise pair during each of the most recent 2 blocks, and the overall percentage correct over the course of those two blocks was at least 75%. For the first 18 subjects, only non-anchor pairs were considered in these calculations, duplicating the methods of Ellenbogen et al. (2007). For later subjects, all premise pairs were taken into consideration when calculating whether to exit training. See Discussion for the rationale for this change. Although the later version of the task is more stringent in the criteria to exit training—and in some cases subjects experienced additional blocks of training that they would not have with the earlier version—the mean number of blocks to reach criteria on the first round of training did not differ between subjects run with the two versions (older version: 7.8±0.9; newer version: 7.2±0.9; $t(33)=0.49$, $p=0.62$). This comparison includes all subjects in Experiment 1, regardless of whether they met inclusion criteria, except for two outlier subjects who completed 30 blocks of training without meeting exit criteria. Regardless of

whether a subject reached these criteria, training would not continue past a total of 30 blocks across all rounds of training.

Following each round of training, subjects were instructed to relax and read magazines for 5 minutes, then had an Immediate Test. Testing was similar to training, except without feedback: after the subject's response, the stimuli merely disappeared, rather than moving to potentially reveal a smiley face. The Immediate Test consisted of 5 blocks of 10 trials, each block consisting of two presentations of each of the 5 premise pairs (once each with the preferred item on the left and on the right), in a different pseudo-random order in each block. To pass the Immediate Test, subjects had to get each inner (non-anchor) premise pair correct at least 50% of the time, and a score of at least 70% overall on the inner premise pairs. The first 18 subjects were allowed only one round of training, and thus only one chance to pass the Immediate Test; but the remaining subjects, after each failed attempt at the Immediate Test, had another round of training and another opportunity to attempt the test, until they had passed the Immediate Test or completed a total of 30 blocks of training. Subjects who did not pass their final attempt on the Immediate Test (n=9) were excluded from further analysis.

Inference Testing consisted of 5 blocks of 18 trials each: 10 trials testing the premise pairs, 6 trials testing the inference pairs (BD, CE, BE), and 2 trials testing pair AF. *Novel-item testing* consisted of 4 blocks of 36 trials each, in randomized order: 10 trials of the premise pairs; 24 trials of the novel item pairs (pairing each of the 6 item in the hierarchy with each of 2 novel items X and Y); and 2 trials of the novel items X and Y against each other. For some subjects, the transition from Inference Test to Novel-item Test was not marked by any pause or new instructions. But for 5 subjects in the 3hr group

and 5 subjects in the 20min/3hr group, these instructions appeared on screen before the novel-item test:

> We have reached the last phase of this session. You will again see abstract patterns. Some of the patterns will be ones you have seen before, and some will be new. The smiley face hides behind a new pattern about half the time. But we still are not showing you where the smiley face is! Just make your best guess, based on what you learned so far. Try to use your gut feeling and respond quickly, don't spend much time thinking about your choices.

Assignment to Groups: Subjects were assigned to groups by the Multi-dimensional Prospective Randomizer Server (PR). The goal of the PR is to keep the groups as closely matched as possible in the characteristics that seem most likely to confound the outcome measure that is to be compared between groups. In this case we were testing the hypothesis that subjects' scores on TI pairs at the 3 hour test would depend on whether they were tested at 20 minutes. We can control this factor—the intended independent variable—but we cannot sufficiently control other factors which could contribute to TI performance at 3 hours. Plausibly, both amount of training exposure, and the level of mastery of the premise pairs, could impact TI performance. Amount of training exposure was constrained to between 4 and 30 blocks; to constrain this more tightly than that would increase the variance in the level of mastery of the premise pairs. If, by random assignment, one group had more training exposure, or higher premise pair performance, than other groups, this would make the interpretation of any differences in TI score at 3 hours ambiguous. Use of the PR was designed to avoid this scenario.

The PR was run as a server on a separate computer so that it could handle simultaneous requests from multiple clients, in case there was more than one subject doing the task on the same day. On the PR computer, records were maintained of all the

previously run subjects who had successfully completed the protocol, with their group assignments, the number of training blocks each one had completed, and their scores on the Immediate Test. The task computer connected to the PR via a TCP/IP socket. As soon as the subject completed their final iteration of the Immediate Test, the task computer submitted their immediate test score and the number of training blocks they had completed (along with the subject ID) to the PR. Subjects were not assigned to a group until the moment that the groups' protocols diverged: at the end of the 20 minute rest period after the Immediate Test, at which point the computer would administer the Inference Test, the Inference Test and Novel-item Test, or no test, depending on group assignment. At the moment in time when the task computer had to decide which test(s) to administer, it would contact the PR to obtain the group assignment. The PR would choose a group for the subject based on the groups' existing composition.

The PR is designed to add subjects to each group at the same rate, so that groups are closely matched in their distribution across time. (Amongst groups that are still accepting subjects, that is; we closed the 20min*/3hr group after running only a few subjects, due to scheduling concerns.) Thus, when one open group has fewer subjects already assigned to it than the other(s), PR assigns the new subject to the smaller group regardless of scores. However, when PR has to choose between two or more groups of equal size, it makes the decision in a more interesting way. When the PR has to make such a decision, it calculates the means and standard deviations of each of the variables of interest, for the entire sample of subjects it knows about—including those already assigned to a group, and subjects who had scores submitted that day but were not yet assigned a group. It then normalizes the values for each variable for each subject by

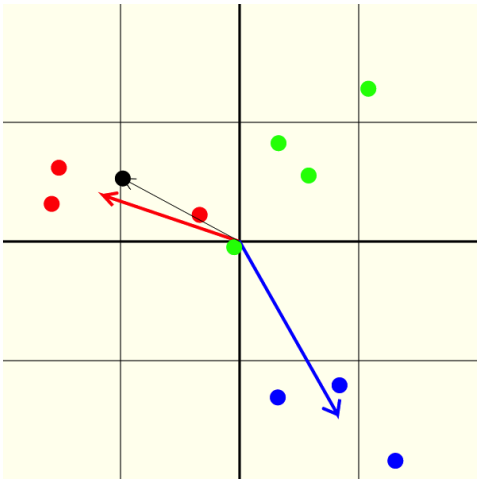

Figure 3.3: Graphical representation of Prospective Randomizer algorithm for assigning subjects to groups. Red, blue, and green dots represent subjects already assigned to groups A, B, and C respectively. The black dot is a subject who has not yet been assigned to a group. X and Y coordinates of each dot are the number of training blocks and premise pair score, each normalized by the mean and standard deviation of the corresponding variable. PR will not consider assigning to group C because it has more subjects than the other two groups. In deciding between groups A and B, it finds the mean vectors for these groups (fat arrows). The dot product of the vector for the new subject (thin black arrow) and each of the group average vectors is calculated. The dot product with the B vector (-1.57) is more negative than the dot product with the A vector (1.33), thus the subject will be assigned to group B. Intuitively this makes sense: the existing composition of group A over-represents, and B under-represents, subjects with parameters similar to the unassigned subject.

subtracting the relevant mean, and dividing by the relevant standard deviation. We can think of each subject as a vector in multidimensional space (2-dimensional, in this case); the origin of the space is the mean of all known subjects, and the length and direction of the vector represents how a subject differs from the mean. The PR then finds the average vector for each group, and the dot product of the new subject's vector with each group average vector, then chooses the group for which this dot product is smallest (or most negative). This has the effect of assigning a subject to that group that most needs to be pulled in that subject's direction; for example, if the groups diverge in mean number of training blocks, and a subject has a high number of training blocks, assigning the subject to the group with the lower mean number of training blocks pulls that group closer to the mean in that dimension. See Figure 3.3.

Protocol: After completing consent

forms, subjects filled out sleep logs and other questionnaires regarding sleep habits. Subjects began the training session of the TI task at approximately 12:00 noon (mean time: 12:09, range: 11:08 to 13:08). Groups did not differ in the mean clock times of the start of training (means ± s.e.m.: 20min/3hr: 12:17 ± 9 min; 3hr: 12:01 ± 9 min; 20min*/3hr: 12:01 ± 1 min; $F(2,25)=0.93$, $p=0.41$), nor in the amount of time they took to train to criteria, defined as the time from their first training trial to the completion of their final attempt at the immediate test (20min/3hr: 8 ± 2 minutes; 3hr: 8 ± 1 minutes; 20min*/3hr: 5 ± 1 minutes; $F(2,25)= 0.60$, $p=0.56$). After the completion of training and the final immediate test, subjects were instructed to read magazines for 20 minutes. At the end of the 20 minute break, the computer connected to the Prospective Randomization Server to assign the subject to one of 3 groups, and administered any test(s) for that group. Subjects in the 20min/3hr group (n=12) did the Inference testing; subjects in the 20min*/3hr group (n=4) did both the Inference testing and Novel-item Test; and subjects in the 3hr group (n=12) did neither test. At the completion of the appropriate test(s), "Thank you! Done for now." appeared on the screen.

Subjects were then given lunch, and allowed to pursue whatever daytime waking activity they wished until they had to return. Most subjects studied or did homework, although some went home or shopping, or ran errands. Subjects returned to the lab for a second visit scheduled to start 3 hours after the start of their first visit. Upon returning to the lab, subjects did a questionnaire including the Stanford Sleepiness Scale (Hoddes, Zarcone, Smythe, Phillips, & Dement, 1973) and questions about their alertness, and then began the 3 hour test. Subjects started the "3 hour test" session, on average, 169 minutes after they started training (range: 154 to 177 minutes). These groups did not differ in this

(means as hours:minutes ± s.e.m. in minutes: 20min/3hr: 2:49 ± 2; 3hr: 2:49 ± 1; 20min*/3hr: 2:51 ± 2; $F(2,25)=0.32$, $p=0.73$), or the length of time interval from the end of training until the 3 hour test (means as hours:minutes ± s.e.m. in minutes: 20min/3hr: 2:39 ± 2; 3hr: 2:39 ± 2; 20min*/3hr: 2:45 ± 3; $F(2,25)= 1.04$, $p=0.37$). For all subjects the 3 hour test consisted of the Inference Test followed by the Novel-item Test.

After subjects completed testing, they completed a debriefing questionnaire shown in Box 3.1 to assess their awareness of the task structure and their strategies. Subjects had to complete each page of the questionnaire before going to the next page, and were not permitted to go back to earlier pages to revise answers. This questionnaire was based on the questionnaire used by (Frank et al., 2005) but modified to elicit information about how subjects encoded the premise pairs, and whether they were aware of, and how they approached, the novel-item pairings. Unlike Frank et al., our awareness questionnaire did not mention the word “hierarchy” until the last page.

Box 3.1

Awareness Questionnaire

Page 1:
Did you have the impression that some of the pairs of patterns were easier to choose between than others? (Yes/No)
Did you think any of the patterns were ALWAYS correct (no matter what the other pattern was)? (Yes/No)
Did you think that any of the patterns were ALWAYS incorrect (no matter what the other pattern was)? (Yes/No)
Did you have any tricks for memorizing the individual patterns or the pairs of patterns? (Yes/No)
If so, explain briefly.

Page 2:
Did you give the patterns names? (Yes/No)
If so, give examples.
Did you have the impression that there was some kind of logical rule or order? (Yes/No)
If so, please explain briefly.

Page 3:
In the test phase, did you notice any new combinations of patterns taken from those you saw before in the training phase? (Yes/No)
How did you make your choice in these cases? (e.g., guessed, went with instinct, used some sort of rule. Please explain.)
In the test phase, did you notice any new patterns that you hadn't see in the training phase? (Yes/No)
How did you make your choice in these cases?

Page 4:
Did you think that there was a hierarchy among the patterns seen in training? That is, did you think they could be ranked from "best" to "worst"? (Yes/No)

Analyses: Data files output by the task program were processed to extract summary information using scripts written in Python version 2.6.1. Statistical tests were performed using R, version 3.2.1. All t-tests were preceded by Bartlett test of homogeneity of variances, and whether the pooled variance was used to estimate the variance, or whether the Welch approximation of degrees of freedom was used, depended on whether the variances were significantly different between groups. To assess whether experiencing

the Inference test at 20 minutes affected inference performance at the 3 hour test, we used a two-tailed t-test to test whether the performance of subjects on transitive inference pairs (BD, CE, and BE) in the 20min/3hr group at their 3 hour test was different from the performance of subjects in the 3hr group at their 3-hour test (at 0.05 level of significance). As this revealed no significant difference, within-group repeated measures of just the 20min/3hr and 20min*/3hr groups were used for all remaining analyses.

To determine whether transitive inference ability develops over the course of approximately 2.5 to 3 hours awake, we used paired t-tests to compare subjects' scores at 20 minutes and 3 hours on transitive pairs, both 1º (BD and CE) and 2º (BE), and on the non-transitive end-anchor pair (AF). Not being sure whether exposure to the novel-item test at the 20 minute test would affect any off-line processes, these comparisons were run on data from the 20min/3hr and the 20min*/3hr group pooled together, and again on just the 20min/3hr group.

Experiment 2

Experiment 2 was designed to determine whether transitive inference ability develops more over the course of a nap than over an equivalent time spent awake, and whether there are correlations between amount of time spent in particular sleep stages and subsequent transitive inference. Subjects trained to criteria on the premise pairs of the six-item hierarchy as used in Experiment 1. Subjects were tested on transitive pairs both at 20 minutes and at 3 hours. All subjects had EEG electrodes attached to their heads. Some subjects (the Nap group) took a PSG-monitored nap between the testing sessions while others (the Wake group) remained awake. See Figure 3.4 for the protocol timeline.

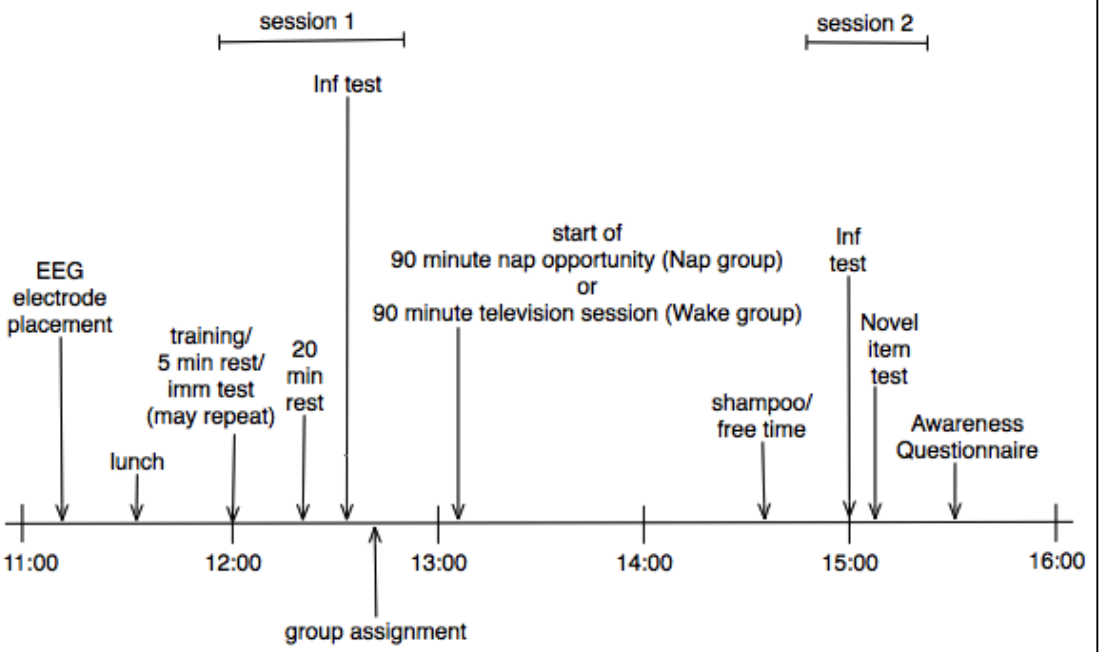

Figure 3.4: Timeline of protocol for Experiment 2.

Participants: A total of 45 subjects (35 female) between the ages of 18 and 30 (mean 22.2, standard deviation 2.8) were recruited via the same recruitment channels as in Experiment 1. In addition to the inclusion and exclusion criteria listed for Experiment 1, subjects were also excluded if they reported drinking more than 5 alcoholic drinks per week, more than 3 caffeinated drinks per day, or working a shift work job. Subjects who reported travel across more than one time zone within the previous month had their participation delayed until at least a month had elapsed. Subjects who did not pass their final attempt on the immediate test (n=8) were excluded from further analysis, and additionally one subject withdrew participation due to illness; thus we report data only from the 36 subjects who met inclusion criteria and completed the protocol. As subjects who have very high scores after only 20 minutes cannot show much later improvement, subjects were recruited until there were 12 subjects in each group who were not at ceiling

on transitive inference pairs at the 20 minute test; and many of the analyses include only this subset of 24 “non-ceiling” subjects.

Task: The task was as described in Experiment 1 above, except that for 4 subjects who had not passed the immediate test by the 21st block of training, blocked-order mode training was introduced, and the maximum number of training blocks was increased to 36. In blocked-order training, blocks 21 and 22 consisted of 4 consecutive presentations of each premise pair; blocks 25 to 27 consisted of 6 consecutive presentations of each pair; and blocks 30 to 33 consisted of 8 consecutive presentations of each pair. (All other blocks were in randomized order.) Within the blocked-order training, each series of trials featuring a particular premise pair would be followed by trials of a non-overlapping premise pair; for example, a series of consecutive presentations of BC might be followed by a series of consecutive presentations of DE or EF, never by AB or CD. To exit a round of training, subjects had to achieve criteria (a score of 75% or better overall, and at least 50% on each premise pair in each block) over the course of two consecutive randomized-order training blocks. Thus, for example, training would not discontinue after block 21, 22, or 23, but could discontinue after block 24. The revised criteria for exiting training (requiring 50% per pair per block, and 75% overall, on *all* premise pairs, not just the internal pairs) were used for all subjects. All subjects performed the Inference test 20 minutes after completing training, and the Inference test followed by the Novel-item test 3 hours after training. All subjects saw the Novel-item instructions before the Novel-item test.

Protocol: Subjects arrived at the lab at 11:00 am. After giving consent, subjects filled out, in REDCap, a sleep log for the prior three nights, the Epworth Sleepiness Scale

(Johns, 1991), the Stanford Sleepiness Scale (Hoddes et al., 1973), and other questions about their sleep habits and alertness level. Electrodes were then attached as described below, then subjects ate lunch. Training on the premise pairs began shortly after noon (mean start of training time amongst non-ceiling subjects: Nap: 12:25 ± 7 minutes; Wake: 12:18 ± 3 minutes; these do not differ, $t(22)= 0.82$, $p=0.42$). The non-ceiling subjects in the Nap and Wake groups did not differ in the length of time from start of training until the end of training (in minutes: Nap: 11 ± 3; Wake: 15 ± 3; $t(22)= 1.01$, $p=0.32$). During the 5 minute rest period(s) between training and Immediate test, and the 20 minute rest period between Immediate test and Inference test, subjects were instructed to read magazines. A pile of magazines on a variety of subject matter and covering a range of reading levels was provided.

At the end of the 20 minute Inference test, the task computer submitted the overall score on inference pairs to the Multi-dimensional Prospective Randomizer server, which assigned subjects to Nap or Wake. For Nap group subjects (n=19), electrode placement was finalized, all light sources were blocked, and subjects were put to bed to try to sleep, and PSG recording was started. Wake group subjects (n=17) were taken on a walking tour of the surrounding hallways and then offered a selection of television shows to watch. Ninety minutes later, Nap subjects were awoken, electrodes were removed from all subjects, and subjects were given the opportunity to wash their hair. Subjects were given the opportunity to read, continue watching television, or walk, until approximately 3 hours had elapsed since the start of training.

Subjects then answered a short REDCap survey about their alertness level and began the 3 hour testing session, consisting of the Inference test followed by the Novel-

item test. Time elapsed between the start of training and the start of the 3 hour test was not different for the Nap and Wake group (for non-ceiling subjects: Nap: 3:07 ± 2 minutes; Wake: 3:05 ± 3 minutes; $t(22)= 0.51$, $p=0.62$). Nor did time between the end of training and the start of the 3 hour test differ (Nap: 2:54 ± 3 minutes; Wake: 2:48 ± 4 minutes; $t(22)= 1.42$, $p=0.17$). Immediately after testing, subjects filled out the Awareness questionnaire.

Polysomnography: A subset of the 10/20 system electrode placement locations were used: scalp locations F3, F4, C3, C4, O1, and O2; mastoid placements A1 and A2; ocular channels LOC and ROC; and chin placements EMG1 and EMG2. The ground electrode was placed on the subject's collarbone and the reference electrode on the subject's forehead. Twelve of these 14 electrodes—all electrodes except those on the chin—were placed before the TI training session. The remaining electrodes, EMG1 and EMG2, were not placed until just before the nap, and only for subjects in the Nap group. Once these electrodes were placed, impedances were checked, and electrodes replaced as needed, to try to get all impedances below 10 kOhm.

Subjects in the Nap group slept in bedrooms equipped with Grass AURA LTM64 polysomnography (PSG) recording systems (Grass Technologies, Rhode Island). Each EEG and ocular channel was referenced against its contralateral mastoid, and the EMG channel consisted of the two chin electrodes referenced against each other. PSG data were visually scored in 30-s epochs according to the sleep staging criteria described in the 2007 version of the American Academy of Sleep Medicine manual for scoring sleep (Iber, Ancoli-Israel, Chesson, & Quan, 2007).

Analyses: Subjects who, at the 20 minute test, were near ceiling on TI ability (having an average score of 90% or above on all the transitive pairs BD, CE, and BE, and a score of 85% or above both on the 1º pairs (BD and CE) and on the 2º pair (BE)) were excluded from between-group comparisons of improvement on the task and sleep stage correlations. Amongst the remaining Nap (n=12) and Wake (n=12) subjects, to determine whether a nap affects transitive inference ability, we did t-tests to compare the mean change in scores from 20 minutes to 3 hours on 1º and 2º pairs between the two groups. Furthermore, amongst the Nap subjects, to assess the relationships between each sleep stage and changes in their transitive inference ability, we linearly regressed the amount of improvement on 1º and 2º pairs against time in REM, N1 sleep, N2 sleep, and N3 sleep.

Analyses on Pooled Experiments 1 and 2 data: Free-text responses on the Awareness Questionnaire were scored independently by 3 judges on a scale from 0 to 3 for the extent to which each subject reported thinking of the stimuli as having a hierarchy from most powerful to least powerful. The median of the three judges' scores for each participant was taken as their "awareness score". Four of the questions on the Awareness Questionnaire, the second and third question on page 1 and the first and third questions on page 3, were yes/no questions regarding whether subjects had noticed and observed objectively true features of the task.

The Novel-item Test was analyzed by fitting a line, using linear regression, to each subject's data individually. (As only 4 subjects did the Novel-item Test at the 20 minute testing session, only data from the 3 hour testing session was included in these analyses.) Points to which a subject's line was fitted had, as their y values, the percentage trials on which each item was chosen when presented with a novel item, and as their x values, the

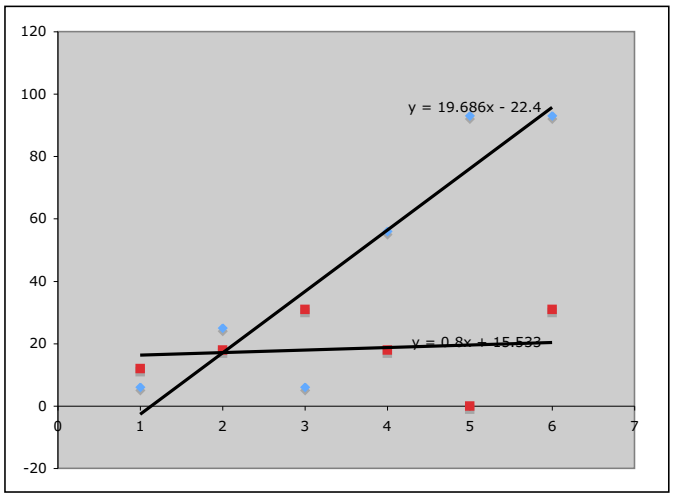

Figure 3.5: Slopes of novel-item scores versus hierarchy position for two example subjects. One subject (blue dots) chooses the hierarchy member over the novel item at a higher rate for higher-ranked items, and the slope of the regression line is about 20 percentage points per hierarchy position. Another (red dots) chooses the novel item at about the same rate when paired with any item of the hierarchy, and the slope of the regression line is close to 0.

items' positions in the hierarchy. The slopes of these lines were taken as the 6-item Novel-item Test (NiT) slopes. (Positions in the hierarchy were coded such that higher-ranked items were numerically higher, so that a positive slope would reflect a tendency to choose higher-ranked items more frequently than lower-ranked items. See Figure 3.5) To test the whether subjects had developed a preference for higher-ranked items in the hierarchy—expressed as a tendency to choose higher-ranked items more frequently than lower-ranked items when paired with novel items—we tested the hypothesis that the mean of the 6-item NiT slopes was significantly greater than zero.

A preference gradient can only support transitive inference performance if the items in the inference probe trials (items B, C, D, and E) are sufficiently distinguished. To test whether subjects had developed preferences for the higher-ranked items amongst the non-anchor items, we fitted a line for each subject to just items B to E on the Novel-item test percentages, and tested whether the mean slope for these lines (the mean 4-item NiT slope) was significantly greater than zero. To test whether preferences gradients, as

revealed on the Novel-item test, supported transitive inference performance, we tested whether subjects with stronger gradients (more positive slopes) had higher scores on the transitive pairs on the Inference Test that immediately preceded the Novel-item Test. To test the hypothesis that a preference gradient based strategy produced more accurate performance on 2º pairs than 1º pairs (the symbolic distance effect), we tested for a correlation between 4-item NiT slope and the difference between 2º score and 1º score—again using data from the same testing session as the Novel-item Test, in case the strength of the preference gradient (and the dominant strategy) changed over time. To test our hypothesis that a stronger preference gradient at the later time results in 2º pairs improving relative to 1º pairs, we regressed change from 20 minutes to 3 hours in 2º score minus 1º score against 4-item NiT slope.

Chapter IV

Results

Experiment 1: One-way ANOVAs were conducted to confirm that all three groups were very similar in their training performance. The three groups did not differ in number of training blocks completed (20min/3hr: 8.8 ± 1.4; 3hr: 8.3 ± 0.9; 20min*/3hr: 8.3 ± 1.8; $F(2,25)= 0.05$, $p=0.95$), their score on premise pairs in the Immediate Test (20min/3hr: 93 ± 2; 3hr: 94 ± 2; 20min*/3hr: 93 ± 2; $F(2,25)= 0.02$, $p=0.98$), or number of iterations of the Immediate Test (20min/3hr: 1.3 ± 0.2; 3hr: 1.3 ± 0.1; 20min*/3hr: 1.0 ± 0.0; $F(2,25)= 0.65$, $p=0.53$). The freakishly high p-values for the tests for numbers of training blocks and Immediate Test scores confirm that PR was successful in assigning subjects to groups strategically to keep the values of these variables well-matched between groups. Random assignment would have had only a 0.1% chance of resulting in groups that were this similar in these parameters. Thus, we were successful in avoiding differences in training history between the groups that could confound the interpretation of any differences found at the 3 hour time.

Next we asked whether the set of tests experienced 20 minutes after training had any measurable effect on subjects' performance when they were tested 3 hours later. Subjects in the 3hr and 20min/3hr groups did not differ in their overall performance on either premise pairs (20min/3hr: 87% ± 3; 3hr: 90% ± 3; $t(22)= 0.64$; $p=0.53$) or transitive inference pairs (20min/3hr: 54% ± 10; 3hr: 50% ± 11; $t(22)= 0.31$, $p=0.76$) at 3 hours. Nor, within the inference pairs, did they differ in their performance specifically on the 1º pairs (20min/3hr: 55% ± 10; 3hr: 50% ± 10; $t(22)= 0.35$, $p=0.73$) or 2º pairs

(20min/3hr: 53% ± 13; 3hr: 50% ± 13; t(22)=0.18, p=0.86). End-anchor pair performance also did not differ (20min/3hr: 88% ± 9; 3hr: 77% ± 10; t(22)=0.79, p=0.44). See Figure 4.1. Thus, having done the inference test at 20 minutes does not appear to affect later performance, and therefore we used within-subject comparisons wherever possible. An insufficient number of subjects were run in the 20min*/3hr group to formally test

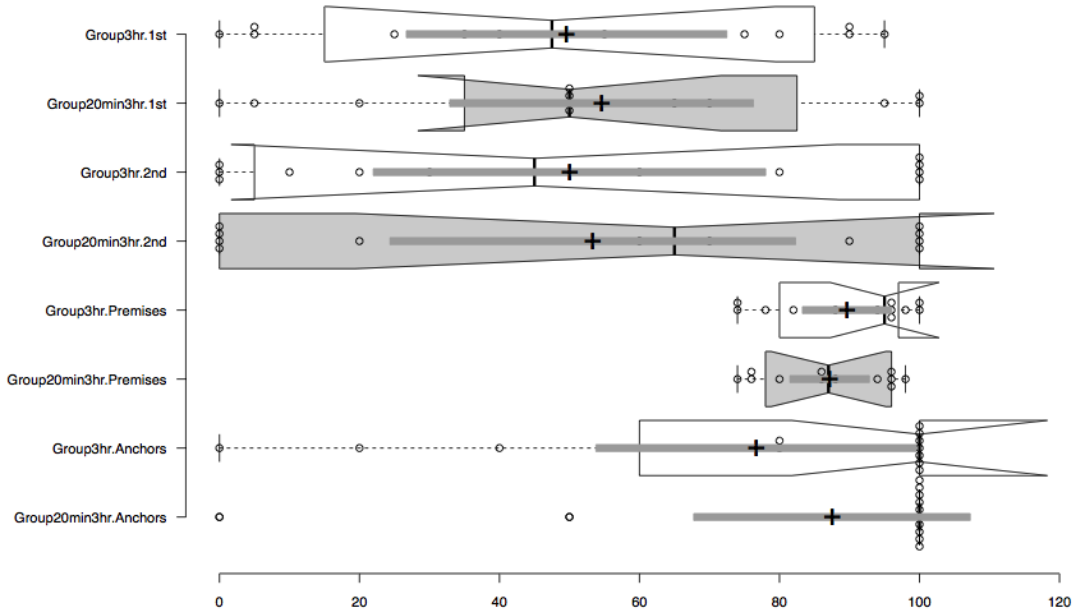

Figure 4.1: Experiment 1, 3 hour comparison. Comparisons of distributions of scores on the 3 hour test between subjects who did (light grey) or did not (white) do the 20 minute Inference Test. Center lines show the medians; box limits indicate the 25th and 75th percentiles as determined by R software; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles, outliers are represented by dots; crosses represent sample means; medium grey bars indicate 95% confidence intervals of the means; data points are plotted as open circles. n = 12 sample points. Plots generated using BoxPlotR at http://boxplot.tyerslab.com/ (Spitzer M., 2014)

whether their performance at 3 hours matched the other two groups, thus we draw no conclusion regarding whether doing the Novel-item Test affects later performance.

We evaluated whether performance on transitive inference pairs changed between the two testing sessions. Paired two-tailed t-tests were conducted to compare 20 minute

and 3 hour scores on transitive inference pairs changed amongst subjects who performed the inference test at both times. (See Table 1.) Pooling results from the 20min*/3hr and 20min/3hr groups, we find no change in score on either 1º pairs (mean ± s.e.m. change in score: +3.8% ± 3.3; t(15)= 1.13, p=0.27) nor on 2º pairs (-1.9% ± 7.2; t(15)= 0.26, p=0.80). Nor did these subjects improve on the end-anchor pair (-3.1% ± 2.2; t(15)= 1.43, p=0.17). Excluding subjects who performed the novel-item test at 20 minutes does not meaningfully change these results for either 1º pairs (+3.3% ± 4.1; t(11)= 0.80, p=0.44), for 2º pairs (-5.8% ± 9.2; t(11)= 0.64, p=0.54), or for the end-anchor pair (-4.2% ± 2.8; t(11)=1.45, p=0.18). In any case, there is no improvement.

This was not because subjects were already displaying transitive inference ability at 20 minutes. Amongst all subjects who achieved an adequate score on the Immediate Test and did the Inference Test at 20 minutes, the mean score on 1º pairs was 55% ± 8, not significantly different from chance (t(15)= 0.59, p=0.56), and the mean score on 2º pairs was 59% ± 10, not significantly different from chance (t(15)= 0.85, p=0.41). These results are consistent with Ellenbogen et al.'s findings in their 20-minute group.[8] With

Table 1
Experiment 1, TI scores for each group, at each testing session (means ± s.e.m.); and changes in score. Change in score is 3 hour score minus 20 minute score, not percentage of baseline.

group	1º at 20 minutes	1º at 3 hours	Δ 1º	2º at 20 minutes	2º at 3 hours	Δ 2º
3hr		50% ± 10			50% ± 13	
20min/3hr	51% ± 9	55% ± 10	3% ± 4	59% ± 12	53% ± 13	-6% ± 9
20min*/3hr	65% ± 18	70% ± 15	5% ± 5	58% ± 21	68% ± 20	10% ± 7

[8] However, our subjects were significantly better than Ellenbogen et al.'s on the end-anchor pair at 20 minutes (our study: 94% ± 5; Ellenbogen et al.: 69% ± 9; t(26)=2.7, p=0.013).

these scores, subjects were overall nowhere near ceiling, and improvements should have been possible. But 3 hour scores on transitive pairs were still not different from chance (all p>0.4).

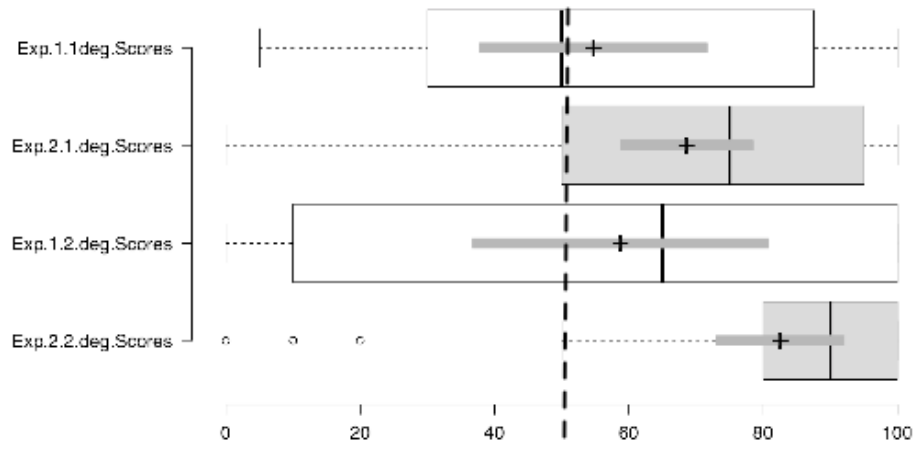

Figure 4.2: Discrepancy at baseline between experiments. Scores on 1º pairs (top) and 2º pairs (bottom) at the 20 minute test, Experiment 1 (white) vs. Experiment 2. Center lines show the medians; box limits indicate the 25th and 75th percentiles as determined by R software; whiskers extend 1.5 times the interquartile range from the 25th and 75th percentiles, outliers are represented by dots; crosses represent sample means; medium grey bars indicate 95% confidence intervals of the means. The heavy dashed line is at 50%, at-chance level of performance. Plots generated using BoxPlotR at http://boxplot.tyerslab.com/ (Spitzer M., 2014)

Experiment 2: Unlike in Experiment 1, transitive inference scores of subjects in Experiment 2 were significantly above chance 20 minutes after completion of the immediate test. At 20 minutes, the mean score on 1º pairs was 69% ± 5 (significantly different from 50%: t(35)= 3.83, p=0.0005), and on the 2º pairs 83% ± 5 (significantly different from 50%: t(35)= 6.89, p<0.0001). (See Figure 4.2.) A paired t-test shows that at 20 minutes, 2º scores are significantly higher than 1º scores (t(35)= 2.9, p=0.007), contrary to what we would expect based on Ellenbogen *et al.*'s results. Whereas in Experiment 1, only 12.5% of subjects who performed the Inference Test at 20 minutes

met our definition of "ceiling" at that point, of the Experiment 2 subjects, 33.3% were at ceiling at 20 minutes. We hypothesized that improvements in TI performance over time were possible amongst subjects who were not already proficient at 20 minutes; for that reason, all analyses below looking at improvements in performance on transitive pairs include only the 24 non-ceiling subjects.

T-tests were conducted to confirm that non-ceiling subjects in the Nap and Wake groups were similar in their training performance and initial (20-minute) performance on transitive inference pairs. They did not differ in number of training blocks completed (Nap: 14 ± 3; Wake: 18 ± 3; $t(22)= 1.08$, $p=0.29$), their score on premise pairs in their final Immediate Test (Nap: 95% ± 2; Wake: 94% ± 2; $t(22)= 0.23$, $p=0.82$), or the number of times they took the Immediate Test (Nap: 1.3 ± 0.1; Wake: 1.4 ± 0.1; $t(22)= 0.41$, $p=0.69$). On the Inference Test at 20 minutes, mean scores achieved by non-ceiling subjects were not significantly different between the Nap and Wake groups on 1º pairs (Nap: 56% ± 8; Wake: 53% ± 7; $t(22)= 0.31$, $p=0.76$) and on 2º pairs (Nap: 73% ± 10; Wake: 76% ± 9; $t(22)= 0.25$, $p=0.80$). Thus, were there any differences in improvement evident at 3 hours, the interpretation of these differences would not be confounded by baseline differences.

We asked whether either the nap or the time elapsed between testing sessions produced any changes in performance on TI pairs. Subjects in the Wake group who were not at ceiling at 20 minutes improved by 2 ± 4 percentage points on 1º pairs, and decreased by 3 ± 3 percentage points on 2º pairs. Subjects in the Nap group who were not at ceiling at 20 minutes showed 5 ± 3 percentage points improvement on 1º pairs, and 3 ± 4 percentage points improvement on 2º pairs, at 3 hours. None of these improvements

were statistically significant, but the change on 1º pairs in the Nap group shows a trend towards improvement ($t(11)= 1.96$, $p=0.08$). Neither the mean improvements in 1º pairs nor in 2º pairs is significantly different between the Nap and Wake groups (1º: $t(22)= 0.52$, $p=0.61$; 2º: $t(22)= 1.47$, $p=0.16$). (See Figure 4.3.) As in the first experiment, no off-line improvements appeared after only 2.5 to 3 hours; and the addition of a nap in that time did not make any apparent difference.

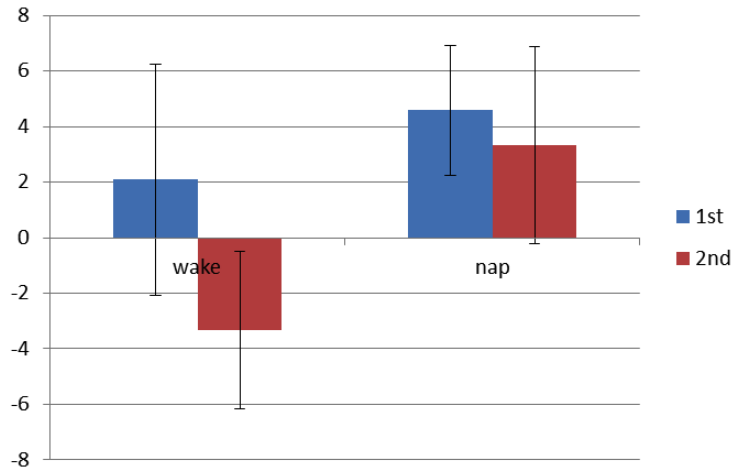

Figure 4.3: Improvements in transitive pair scores for non-ceiling subjects were not significant in either group. (Error bars are s.e.m.)

In spite of the lack of improvement on average, there remained the possibility that differences in improvement amongst subjects were related to sleep parameters. Table 2 shows a summary of nap polysomnography data from the 12 non-ceiling subjects in the Nap group. Linear regressions of improvement on 1º pairs and 2º pairs against total sleep time reveal that total sleep time does not predict either. For 1º pairs the slope of the regression line is -0.03, indicating a 0.03 percentage point decrease in score per 30-second epoch of sleep. This is neither

Table 2
Nap sleep architecture:

N1	5.3 (4.8)
N2	36.0 (9.0)
N3	19.9 (15.3)
R	12.8 (8.0)
Total Sleep Time	74.0 (15.4)

Values are mean (standard deviation) minutes in each sleep stage.

statistically significant (p=0.87) nor meaningfully significant. For 2º pairs the slope is -0.37, again neither statistically significant (p=0.13) nor meaningfully significant. Multiple regression of 1º pair and 2º pair improvements against time in each stage of sleep (N1, N2, N3, and REM) reveal one significant coefficient: scores on 2º pairs decrease by 0.52 percentage point per 30-second epoch of N3 sleep (p=0.048). But the overall p-value of the model was 0.21, indicating a 21% chance that coefficients at least as large as these would be found in the absence of any real relationships. Thus we cannot reject the null hypothesis, which states that sleep architecture had no effect on change in TI scores.

Analyses on Pooled Data: Due to a computer issue, trial-by-trial data from the training portion of the task for one subject in the Nap group was lost. This subject was at ceiling at 20 minutes, and thus not included in the analyses of improvements or sleep correlations above. This subject is included in analyses below for which we have the relevant data.

Awareness Questionnaire: We asked how many of the subjects gave responses on the Awareness Questionnaire that were consistent with the use of explicit logical reasoning strategies. In some studies (Werchan & Gomez, 2013) such subjects would be excluded from the main analyses, and in others (Frank et al., 2005) they would be analyzed separately. We wanted to find out whether there were enough of these "aware" subjects to justify re-doing analyses with the aware subjects excluded, in case they were obscuring group effects. Pooling data across the two experiments, and including subjects who were at ceiling on TI pairs at 20 minutes, only 12.5% of subjects were given a median score of 3 (indicating a high degree of awareness of the hierarchy) on the free-

text responses on the Awareness Questionnaire. Each of these subjects also answered "yes" to the "Did you think that there was a hierarchy…" question. Surprisingly, however, having a high awareness score did not necessarily entail correctly answering yes/no questions to which subjects should have answered "yes" if they successfully observed and explicitly remembered some aspects of the structure of the task. Of the 18 subjects who obtained a score of 2 or 3, 89% answered "yes" to "Did you think any of the patterns were ALWAYS correct (no matter what the other pattern was)?", but only 67% answered "yes" to "Did you think that any of the patterns were ALWAYS incorrect (no matter what the other pattern was)?" and a mere 56% answered "yes" to "In the test phase, did you notice any new combinations of patterns taken from those you saw before in the training phase?" and 56% answered "yes" to "In the test phase, did you notice any new patterns that you hadn't see in the training phase?" Overall there was a correlation of only 0.07 between awareness score and the number of these questions answered correctly amongst the included subjects. This is a very small correlation, and not statistically significant ($p=0.58$). This indicates a surprising dissociation between being able to correctly answer questions about what was experienced during the task, and either noticing or being able to explain that the stimuli could be organized into a hierarchy. Thus, very few subjects gave responses on the Awareness Questionnaire that we would expect from someone using logical reasoning strategies. Only 2 subjects out of the 64 who met inclusion criteria both had a median score of 3 on the free-text responses and answered all 4 yes/no questions with objectively true answers. Both of these subjects were in the Nap group of Experiment 2, and were already excluded from analyses of improvements and sleep stages due to being at ceiling at 20 minutes. This is consistent

with our expectations that subjects who fully understood the task at an explicit level would be at ceiling.

We wanted to see whether responses on the Awareness Questionnaire questions predicted TI performance. To discover which questions (if any) on the Awareness Questionnaire best predicted TI performance, we fitted—for all subjects who met criteria and did the Inference Test at 20 minutes—a linear regression model with 20 minute score on all transitive pairs as the dependent variable, and, as the independent variables, awareness score, whether subjects had responded "yes" to each of the questions with objectively true answers, and whether subjects had answered "yes" to the question regarding whether they thought there was a hierarchy. The overall model was not statistically significant ($F(6, 45)= 1.15$, $p=0.35$) and explained only 13% of the variance in initial TI score. ($R^2= 0.13$, adjusted $R^2= 0.017$) The largest coefficient, by far, was for the question "Did you think that there was a hierarchy among the patterns seen in training? That is, did you think they could be ranked from 'best' to 'worst'?" According to the multiple regression model, factoring out the impact of all answers to other questions, subjects who respond "yes" to this question achieved an initial overall TI score 19 percentage points higher than those who did not. This coefficient is still not significant in the model ($p=0.06$). However, a t-test comparing subjects who answered "yes" to this question versus those that answered "no" yields a significant difference between these groups on initial TI score (answering Yes: 73% ± 4; answering No: 52% ± 5; $t(50)= 2.59$, $p=0.01$). Interestingly, the initial TI score among those answering "yes" is significantly above chance ($t(38)=5.2$, $p<0.00001$) but the initial TI score among those answering "no" is not ($t(12)=0.33$, $p=0.74$). The coefficient of the awareness score was not significant in

the model, but awareness score is highly related to whether subjects answered "yes" to the hierarchy question. Subjects who answered "yes" had a mean awareness score of 1.3 ± 0.2, whereas those answering "no" had a mean awareness score of 0.3 ± 0.1. The difference is highly significant ($t(61.1)=5.4$, $p<0.0001$). After removing the hierarchy question from the model, and scaling the awareness score scale to the same range of possible values as the other factors, the coefficient of the awareness score becomes the largest coefficient in the model (13.9 percentage points TI score per 3 awareness score points), but still not significant ($p=0.25$). A linear regression of initial TI scores against just awareness score is not significant ($F(1,50)=2.0$, $p=0.16$). Thus, only the question "Did you think that there was a hierarchy…" proved to be useful for predicting TI performance.

Novel-item Test: We looked at whether the results of the Novel-item Test, administered immediately after the last Inference test, showed evidence of preference gradients. The mean slope of the regression lines produced for each subject by the Novel-item Test, the 6-item NiT slopes (see Figure 4.4) at the 3 hour testing session, was a highly significant 8.8 ± 1.0 ($t(63)= 9.0$, $p<0.0001$; Figure 4.4). This positive slope indicates that typically subjects choose higher-ranked items over

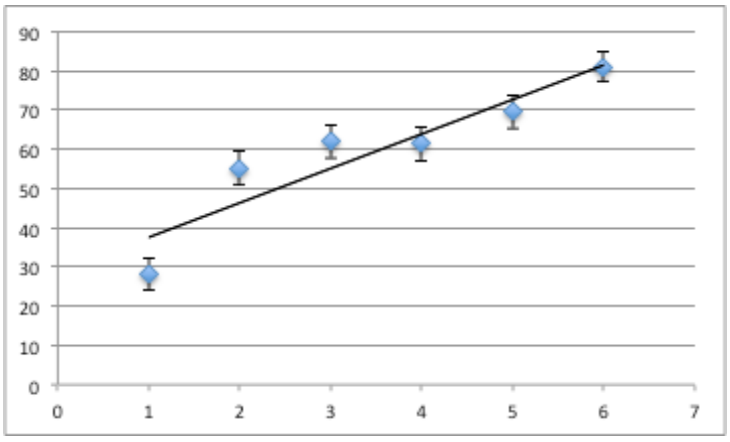

Figure 4.4: Mean 6-item NiT slope: mean rate at which subjects chose each hierarchy item over a novel item, with mean regression line. Item 1 = F and item 6 = A. (Error bars are s.e.m.)

novel items at a higher rate than lower-ranked items. A subject's 6-item NiT slope was correlated with their performance on 1º pairs ($r=0.46$, $p=0.0001$), 2º pairs ($r=0.48$, $p<0.0001$), and overall TI performance ($r=0.52$, $p<0.0001$) at the same session. Thus, we see that subjects are overall more likely to choose higher-ranked items over lower-ranked items in the Novel-item Test, and their tendency to do so predicts their TI performance.

Arguably the significantly positive novel-item slope could be driven entirely by the end items. Since items A and F are rewarded during training 100% and 0% of the time, subjects could have a strong tendency to choose A over the novel item, and anything else over F. This could drag the higher end up and the lower end down without there being a preference gradient across the middle, non-anchor members of the hierarchy. To rule this out we repeated these analyses using only items B through E of the novel-item test (the 4-item NiT slopes). When we did this, the mean slope was somewhat flatter, 4.2 ± 1.5, but still significant ($p=0.006$). Strikingly, the correlations with 2º pair and 1º pair scores became larger (2º score: $r=0.58$, $p<0.0001$; 1º score: $r=0.59$, $p<0.0001$), with correlation with overall TI score even stronger ($r=0.65$, $p<0.0001$). (See figure 4.5 and Table 3.) Thus, the NiT slope *over only the non-anchor hierarchy members* more strongly predicts

	Mean (± s.e.m.) of slopes for all subjects	Correlation with 1º score	Correlation with 2º score	Correlation with overall score
Novel-item slope fitted to A-F	8.8 ± 1.0	$r=0.46$	$r=0.48$	$r=0.52$
Novel-item slope fitted to B-E	4.2 ± 1.5	$r=0.59$	$r=0.58$	$r=0.65$

Table 3: Novel-item slope correlations

transitive inference performance. It better predicts performance than the slope fitted to all points, suggesting that the 6-item slope is, indeed, sometimes spuriously driven by the anchor items.

Then we asked whether evidence of a preference gradient on the Novel-item Test predicts symbolic distance effects (SDE)—a tendency to do better on 2º pairs than 1º pairs. The correlation between 4-item NiT slopes, and SDE as measured by the subtractive difference between 2º scores and 1º scores, was positive ($r = 0.14$) but not significant ($p=0.27$). Inspection of the scatter plot (see Figure 4.6) shows that the distribution is very non-linear: the *magnitude* of the 4-item NiT slope does not correlate with the *magnitude* of the disparity in 2º and 1º scores. However, we can sensibly ask a less daring question. Is there a relationship between being able to do 2º pairs at least as

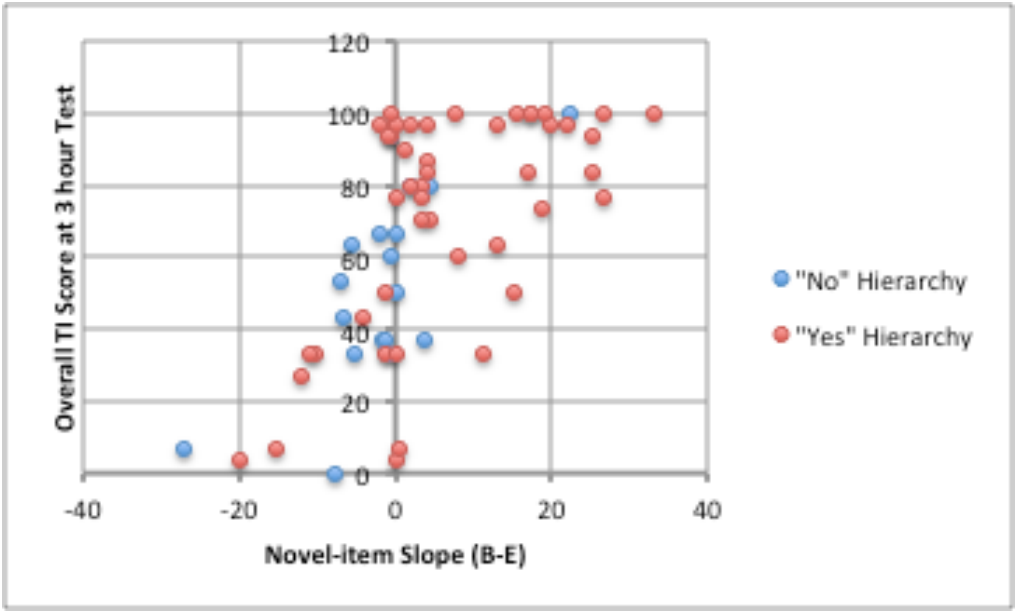

Figure 4.5: Overall scores on inference pairs vs. slopes of lines fitted to scores on items B through E of the Novel-item Test (both measured at the 3 hour testing session), color-coded by whether the subject answered "Yes" to the last question of the Awareness Questionnaire ("Did you think that there was a hierarchy…").

well as 1º pairs (i.e. whether the 2º score – 1º score is at least 0), and having a positive 4-item NiT slope? The results of a chi-square test are highly significant ($\chi^2(1, N=64)=11$, $p=0.0007$), reflecting the fact that 72% of subjects fell into the upper right or lower left quadrant of the plot in Figure 4.6. Thus, in accordance with our expectations, evidence of a preference gradient on the Novel-item Test did strongly predict who would achieve at least as good a score on 2º pairs than 1º pairs. However, subjects with higher NiT slopes showed higher overall performance.

These effects could have perhaps been because subjects who had, or developed, a preference gradient (as measured by the 4-item NiT slope) improved more on 2º pairs over the course of the afternoon. But, subjects overall did not change in their performance on 2º pairs between 20 minutes and 3 hours (improvement amongst all subjects: -0.8% ±

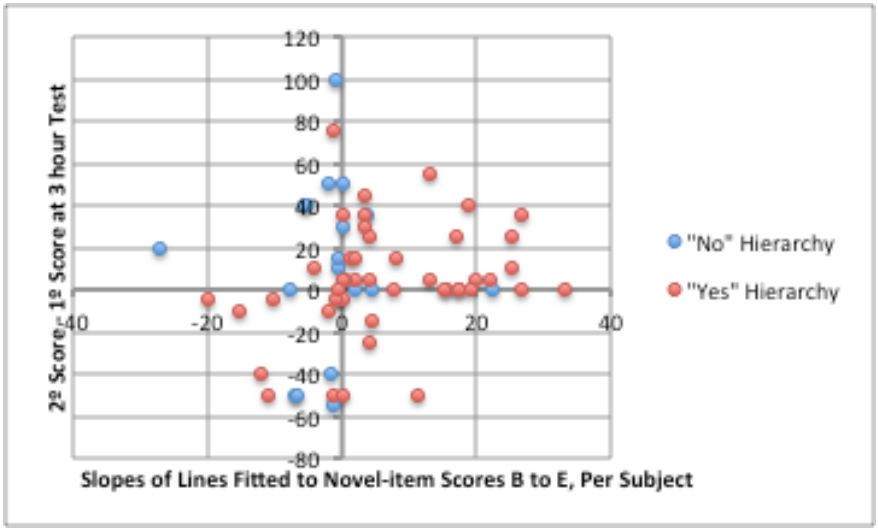

Figure 4.6: Difference in 2º and 1º scores (positive indicates higher 2º performance) vs. slope of line fitted to scores on items B through E of the Novel-item Test, evidence of a preference gradient; color-coded by whether the subject answered "Yes" to the last question of the Awareness Questionnaire ("Did you think that there was a hierarchy…").

2.4; t(51), p=0.75). However this outcome could be the result of those subjects who had a preference gradient showing improvement, while other subjects got worse.

To examine this possibility, we compared the improvement in 2º pairs in subjects having a positive 4-item NiT slope with those who did not. The difference in improvement was not significant (positive 4-item NiT slope: 0.7% ± 1.6; zero or negative 4-item NiT slope: -2.5% ± 4.9; t(50)=0.66, p=0.51). Nor were these groups different in their change in 1º scores (positive 4-item NiT slope: 0.7% ± 2.0; zero or negative 4-item NiT slope: 4.8% ± 2.4; t(50)= 1.3, p=0.20). If we measure symbolic distance effects (SDE) as 2º score minus 1º score, subjects with positive 4-item NiT slopes did not change from 20 minutes to 3 hours in their SDE, whereas other subjects actually *decreased* in SDE—showed poorer performance on 2º pairs relative to 1º pairs at the later time—although this difference was not significant (positive slope: 0.0 ± 2.3; zero or negative slope: -7.3 ± 5.2; t(50)=1.4; p=0.18). Thus, based on analyses in which we pool together all subjects (except those in the 3hr group from Experiment 1, for whom we cannot measure improvement), it does not appear that having a preference gradient—as shown on the Novel-item test at 3 hours—results in overall improvement on TI performance, nor does it produce relative improvement on 2º over 1º pairs.

We asked whether sleep changed these results by re-running these analyses separately on the Nap and Wake groups of Experiment 2. In neither group did subjects improve, on average, on 2º pairs (mean improvement in Nap group: 1.6% ± 2.3; t(17)=0.68, p=0.51; in Wake group: -2.4% ± 2.0; t(16)= 1.2, p=0.26). In the Wake group, change in 2º pair score between the two testing sessions did not differ between subjects who did or did not have a positive 4-item NiT slope (positive slope: -4.2% ± 2.3; zero or

negative slope: 2.0% ± 3.7; t(15) = 1.4, p=0.17). However, improvement on 2º pairs did differ significantly between Nap subjects having a positive 4-item NiT slope and other Nap subjects (positive slope: 5.8% ± 3.3; zero or negative slope: -5.7 ± 3.7; t(18)=2.8, p=0.01). (See Figure 4.7) Improvement on 1º pairs did not differ significantly between subjects who did or did not have a positive 4-item NiT slope in either group (both p>=0.29). In the Wake group, the change in SDE (defined as 2º score minus 1º score) from 20 minutes to 3 hours did not differ between those having a positive 4-item NiT slope and the others (positive slope: -3.3% ± 4.5; zero or negative slope: -5.0% ± 8.4; t(15)=0.19, p=0.85); but, within the Nap group, change in SDE did differ significantly between these subsets (positive slope: 4.6 ± 2.3; zero or negative slope: -10.7 ± 3.5;

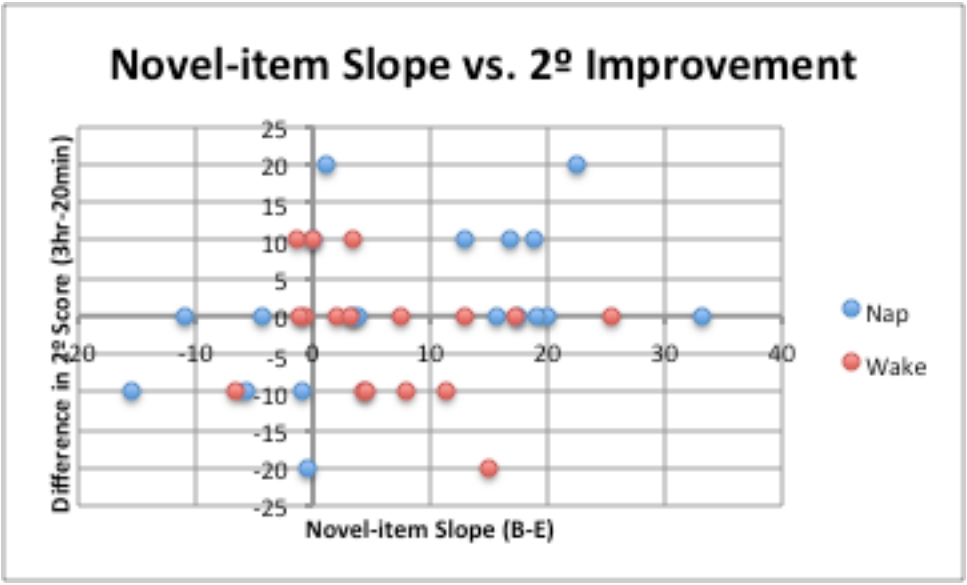

Figure 4.7: Subjects in Experiment 2, Novel-item slope for non-anchor items vs. improvement in 2º scores, color-coded by group. Note that only in the Nap group (blue dots) is there a relationship between NiT slope and direction of change in 2º scores.

t(11)=2.0, p=0.0016). Change in SDE correlated with 4-item NiT slope in the Nap group (r=0.52, p=0.02) but not in the Wake group (r=0.20, p=0.45). (See Figure 4.8.) A repetition of Wake group analyses, in which we also include those subjects from Experiment 1 who did the Inference test twice, produces substantially the same results. (See Table 4.) Thus, subjects show opposite patterns of improvement, depending on whether they showed a preference gradient at the 3 hour testing session as measured by the 4-item NiT slope, *only* if they slept between testing sessions. However, regressing 4-item NiT slope and change in SDE against time in each sleep stage reveals no significant coefficients (all p>=0.2).

Figure 4.8: Subjects in Experiment 2, Novel-item slope for non-anchor items vs. change in SDE, color-coded by group. Only in the Nap group (blue dots) is there a relationship between NiT slope and change in SDE.

Table 4: Changes in TI scores with and without sleep

	Experiment 2 Wake	20min/3hr + 20min*/3hr + Experiment 2 Wake	Experiment 2 Nap
Improvement in 2º scores from 20 minutes to 3 hours different from 0?	-2.4% ± 2.0; t(16)=-1.2, p=0.26	-2.1% ± 3.6; t(32)=-0.59, p=0.56	1.6% ± 2.3; t(17)=0.68, p=0.51
Improvement in 2º scores with positive NiT slope	-4.2% ± 2.3	-3.1% ± 1.8	5.8% ± 3.3
Improvement in 2º scores with non-positive NiT slope	2.0% ± 3.7	-1.2% ± 6.9	-5.7 ± 3.7
t-test of 2º improvements comparing subsets based on NiT slope	t(15) = -1.4, p=0.17	t(31)=-0.27, p=0.79	t(18)=2.8, p=0.01
Improvement in 1º scores with positive NiT slope	-0.8% ± 4.0	0.3% ± 3.1	1.3% ± 2.4
Improvement in 1º scores with non-positive NiT slope	7.0% ± 5.4	4.7% ± 3.4	5.0% ± 1.9
t-test of 1º improvements comparing subsets based on NiT slope	t(15)=-1.1, p=0.29	t(31)=-0.95, p=0.35	t(17)=-1.1, p=0.30
Change in SDE with positive NiT slope	-3.3% ± 4.5	-3.4% ± 3.5	4.6 ± 2.3
Change in SDE with non-positive NiT slope	-5.0% ± 8.4	-5.9% ± 7.2	-10.7 ± 3.5
t-test of changes in SDE comparing subsets based on NiT slope	t(15)=0.19, p=0.85	t(31)=0.30, p=0.77	t(11)=2.0, p=0.0016
Correlation of NiT slope and change in SDE	R=0.20, p=0.45	R=0.13, p=0.46	R=0.52, p=0.02

We asked to what extent performance on the Novel-item Test reflected subjects' awareness that there was a hierarchy. A t-test revealed a significant difference between the mean slope on inner hierarchy members on the Novel-item test between subjects who answered "yes" or "no" on the hierarchy question on the Awareness Questionnaire (Yes: 6.6 ± 1.8; No: -1.9 ± 2.2; t(40.4)=3.1, p=0.004). Thus it appears that when subjects can use a preference gradient, they are aware of it. Based on comparing the Nap and Wake groups of Experiment 2, whether subjects napped did not affect whether they answered

"yes" to the hierarchy question ($\chi^2(1)=0$; p=1), nor did it significantly affect their awareness score (Wake: 1.1 ± 0.3; Nap: 1.5 ± 0.3; t(34)= 1.1, p=0.29).

Having found that both one question on the Awareness Questionnaire and the 4-item NiT slope strongly predicted TI performance, we wondered if these factors mediated the difference in initial scores between Experiment 1 and Experiment 2. The mean 4-item NiT slope was significantly different between Experiment 1 and Experiment 2 (Experiment 1: 0.6 ± 2.3; Experiment 2: 7.0 ± 1.8; t(62)= 2.2, p=0.03). We fitted a multiple regression model with initial overall TI score as the dependent variable, and experiment, 4-item NiT slope, and whether subjects answered "yes" to the hierarchy question as the independent variables. The model was highly significant ($p<0.001$), with one highly significant coefficient, for the 4-item NiT slope: subjects' overall TI increased by 1.4 per unit increase in 4-item NiT slope. Neither the hierarchy question nor Experiment significantly predict TI performance in this model. If 4-item NiT slope is excluded from the model, the hierarchy question's coefficient becomes marginally significant (p=0.044), but this is not surprising given that the hierarchy question so strongly predicts the 4-item NiT slope. With either 4-item NiT slope or the hierarchy question in the model, Experiment was not a significant factor ($p>0.12$). Thus, the difference in 20 minute scores between Experiment 1 and Experiment 2 scores is explicable in terms of the difference in their preference gradients; although this leaves the differences in their preference gradients unexplained.

Chapter V

Discussion

Summary of Results:

• Subjects do not improve on TI pairs over the course of 2.5 to 3 hours spent awake.

• On average, subjects' TI scores—both 1º and 2º—do not change over the course of a 90-minute afternoon nap.

• Whether a subject has a positive 4-item NiT slope predicts the direction of change in their 2º scores over the course of a 90-minute afternoon nap; positive-slope subjects improve, while the others get worse on 2º pairs. The 4-item NiT slope predicts the change in a subject's relative 2º versus 1º performance over the course of a nap.

• Regardless of sleep, subjects with positive 4-item NiT slopes have, at the 3 hour testing session, equal or higher scores on 2º pairs compared to 1º pairs.

• A subject's 4-item NiT slope correlates highly with their TI scores on both 1º and 2º pairs at the same testing session.

• Subjects who answer "yes" to the question "Did you think there was a hierarchy among the patterns seen in training? That is, do you think they could be ranked from 'best' to 'worst'?" have higher Transitive Inference scores than subjects who answer "no", and have higher 4-item NiT slopes.

• Number of minutes in any sleep stage does not predict change in TI performance, change in relative 2º vs. 1º performance, or 4-item NiT slope.

• Subjects' performance on Transitive Inference pairs at 3 hours was unaffected by whether they did an Inference Test at the 20 minute time point; thus doing repeated TI testing appears to not be problematic.

Lack of improvement in TI over a short wake period: Our finding that subjects show, on average, no improvement on the Transitive Inference task at 3 hours in comparison to their performance after only 20 minutes, contrasts with Ellenbogen *et al*.'s finding that as a group, subjects tested on Transitive Inference pairs at least 12 hours after training performed much better than those tested after only 20 minutes. These results contrast, but they do not contradict: these results could be pieces of the same puzzle if the off-line process responsible for the apparent improvement in the Ellenbogen study requires between 3 and 12 hours to produce measurable effects.

The shorter period of off-line time before the final Inference Test is the most obvious difference between the Ellenbogen *et al*. study and the present study, but before concluding definitively that the shorter time period was responsible for the lack of improvement, other reasons for the differing result should be considered. In our version of the task, due to altering the criteria for exiting training, some subjects experienced more training blocks than they would have in the version of the task used in Ellenbogen *et al.*'s study. However, if the two subjects for which this was the case in the 20min/3hr group are excluded, the mean improvement in that group is still less than 5 percentage points for both 1º pairs and 2º pairs. In contrast, Ellenbogen *et al.* reported that subjects who did the inference test after 12 hours awake attained scores that were 23% better on 1º pairs, and 15% better on 2º pairs, on average, than subjects who did the inference test after 20 minutes. Mean improvement on 2º pairs in our 20min/3hr group was actually

negative, driven by one subject who obtained a perfect score on 2º pairs at 20 minutes and 0% at 3 hours. However, if we consider this subject an outlier and calculate the mean improvement on 2º pairs, it is still less than 3%. Theoretically improvement by a subset of subjects could be obscured into non-significance if other subjects are at ceiling at 20 minutes, and therefore cannot improve. However, only one subject in our 20min/3hr group was at ceiling.

Alternate explanations for apparent improvement in prior work: Ellenbogen *et al.* reported differences in TI ability in groups of subjects tested at different times relative to training. As these groups were supposed to be otherwise the same—random samplings of subjects from the same population—this implies that TI abilities of individual subjects evolved over time; *i.e.*, had subjects in the 24 group been tested at 20 minutes, their TI scores would have been lower. However, as in all such studies, the sample is actually a convenience sample, not a random sample. It is ethically and logistically impossible to randomly select from the population and press the randomly-selected into service. Rather, one has to take the subjects who hear about the study, consider it worthwhile, and are willing to come to the lab on the schedule that the study requires; and hope that this set of subjects does not differ from the population in any important way. This introduces the possibility of bias, which could be significant, since the time requirements for different groups were different. One can only speculate, but perhaps the subjects who were willing to continue participation, once they were placed into groups which required more than one visit to the lab (the 12 and 24 hour groups), were the ones who took the task more seriously than those who showed up for the 20-minute protocol. Perhaps the 20-minute group over-represented subjects who had full-time employment (as this was the one

group with no 9 a.m. session). Perhaps this group over-represented subjects with lower executive functioning (as this was the one group that only had to come to the lab once, rather than two visits with the second visit strictly scheduled relative to the first). Given the diversity of strategies used by human subjects on this task, the possibility that such subtle factors could explain at least some difference between groups cannot be ruled out. The marked difference between 20 minute scores in subjects recruited for the two experiments in the current study clearly demonstrates that different cohorts perform differently on this task. Thus, the performance of Ellenbogen *et al*'s 20 minute group *cannot* be taken as a proxy for how their other subjects would have performed at 20 minutes. The differences they report between 20 minutes and the later times may have been group differences. Further studies are required to determine how much within-subject improvement occurs over 12 hours, and whether the within-subject improvement is sufficient to explain Ellenbogen *et al*'s between-group differences.

The change in training exit criteria: A prerequisite to having within-subject improvement is that a sufficient number of subjects start off with a low baseline score when tested soon after training. Ellenbogen et al. (2007) ascribed the apparent development of TI ability in their subjects only after a delay, rather than immediately (as found in other studies), to unique features of their training procedure: specifically, first, the fact that they did not require an extremely high level of mastery of the premise pairs before the end of training; and secondly, the randomized order of the training trials.

> If... inference can take time to develop, then why have previous studies described the ability for inference immediately after training, without the need for such offline delays? Common among these past studies, and distinct from our paradigm, is that participants were trained

> to ceiling levels on the premise pairs. (Ellenbogen et al., 2007)

To avoid training participants to ceiling levels on the premise pairs, training exited as soon as[9] a participant met both these criteria on two consecutive blocks: an average score of at least 75% on the non-anchor pairs, and at least 50% on each non-anchor pair within each block. They justified ignoring participants' performance on the anchor pairs in deciding when to exit training by pointing out that the subjects did not need to master the anchor pairs in order to respond correctly to the inference pairs, as these were all novel pairings between non-anchor items: "The middle pairs were used for criterion, rather than all pairs, because the middle pairs were the building blocks of inference (e.g., one must learn B>C and C>D to answer the inference question: B?D)." (Ellenbogen et al., 2007)

However, these exit criteria are so lax that there is a high probability of meeting them entirely by chance before sufficient learning has occurred. If no learning has occurred, and thus the subject is responding randomly on each trial, the probability of responding correctly at least once on any one premise pair during a single block is $(1-0.5^2) = 0.75$. The probability of responding correctly at least once on each of the three non-anchor pairs for two blocks in a row is $0.75^6 = 18\%$. To simultaneously achieve an overall score of 75% on non-anchor pairs, the subject must respond correctly both times a particular premise pair is presented within a block in at least 3 out of the 6 cases. There is a 0.25 probability of getting any one premise pair within one block correct both times, and a 0.5 probability of getting it correct exactly once; thus, over the course of 2 blocks, the probability of getting both instances of a premise pair within a block correct in n

[9] But not before completing at least 3 training blocks.

cases, and getting exactly one correct in the other (6-n) cases, is ${}_6C_n * (0.25)^n * (0.5)^{(6-n)}$, where ${}_nC_k$ symbolizes the number of ways of selecting k trials out of n to be successes. Thus, the probability of getting both correct in *at least* 3 cases (while getting one correct in the remaining cases) is:

$${}_6C_3 * (0.25)^3 * (0.5)^3 + {}_6C_4 * (0.25)^4 * (0.5)^2 + {}_6C_5 * (0.25)^5 * (0.5)^1 + {}_6C_6 * (0.25)^6$$

This turns out to be only about 0.06.[10] Thus a subject who is responding purely randomly has a 94% chance of continuing training after block 4. However the subject then has a chance of exiting training after block 5. This chance is not 6%, because performance on blocks 3+4 and performance on blocks 4+5 are not independent events; if a subjects is doing block 5, the probability of block 4 having had a 50% score on all three premise pairs is depleted, since all the cases in which blocks 3+4 meet criteria are removed. On blocks 5 and onwards, the chances of the previous block having had a 50% score on all three premise pairs, and 0, 1, 2, or 3 premise pairs at 100%, are reduced by approximately 2%, 11%, 30%, or 42%, respectively. Nonetheless, there is still about a 4% chance of exiting by chance after random guessing after any block from 5 onwards. At this rate, after 10 blocks, only 72% of guessers would still be in training; and survival drops below 50% after only 19 blocks. Of course, if a subject has some partial knowledge leading to above chance on some inner premise pairs (despite still not having sufficient knowledge to pass the subsequent immediate test) the survival curve becomes even more dire. As a result of so many subjects being able to reach criteria purely by chance, many subjects do not pass the immediate test. Subjects without adequate performance on the

[10] The number of ways of choosing k ordered items out of n is n!/(n-k)!. This divided by k!, the number of ways to order k items, gives n!/(n-k)!k!, the number of ways of choosing k trials out of n to be successes.

premise pairs at the Immediate Test were excluded from both Ellenbogen *et al.*'s and our study, as unpublished data from Ellenbogen show that such subjects are (unsurprisingly) rarely above chance at any transitive pair at any time-point.

At the start of this study, for the first 18 subjects, we used Ellenbogen's criteria for exiting training, and subjects were given only one opportunity on the immediate test. Of these 18 subjects, 16 performed fewer than 30 (the maximum number) training trials; they had met the criteria for exiting training. However, 6 (37.5%) of these subjects did not pass the immediate test. Unpublished data show that of the subjects Ellenbogen *et al.* ran, about 40% also did not meet the inclusion criteria at Immediate Test, and were not included in the reported analyses. In retrospect this is not surprising, given the analysis above of the likelihood that a subject responding entirely at chance will exit training before reaching the maximum number of training blocks. Another possible factor is the fact that there may be some rapid forgetting of the premise pairs over the 5 minute rest period, and thus a subject who has an adequate grasp of the premise pairs at the end of training may not remember them as well at the immediate test; especially at a first attempt at the immediate test, as they may not realize at first that they will be tested, and thus have less motivation to remember.

A paradigm that excludes 60% of people who sign up for the study from analysis would seem to produce results of dubious generality. Thus, we altered the task to avoid exiting training prematurely by chance, and to avoid having to exclude subjects who performed poorly on the immediate test. Examining the data from our first 18 subjects revealed that subjects who passed the immediate test usually had very high performance on the end-anchor premise pairs (pairs AB and EF) at the end of their training, whereas

subjects who did not pass the immediate test often had lower performance on the end-anchor premise pairs at the end of training. This is consistent with the theory that many of the ones who failed the immediate test had exited training prematurely by chance. If one is at chance across all the premise pairs, but by luck gets each of the three internal premise pairs right at least once in each of two successive blocks, one has only a 32% chance of achieving the same on both the end-anchor pairs as well. Also having the end-anchor pairs count towards the overall score, which must meet or exceed 75%, decreases the chance of exiting training prematurely as well. The task was also modified to resume training if the subject did not pass the immediate test, to deal with the (still possible) case that subjects would exit training by chance without adequate learning of the inner premise pairs, and the possibility that after training, some subjects may display rapid forgetting of the premise pairs.

Altering the criteria for exiting training based on the assumption that subjects learn all the premise pairs at approximately the same rate—and thus, if they know the inner premise pairs adequately, will probably display high performance on the end-anchor pairs as well—runs the risk of over-training a subject whose learning of the end-anchor pairs trails that of the inner pairs. However, given all the prior work showing a serial position effect in performance on the premise pairs (i.e., the typically better performance on the end-anchor premise pairs compared to inner premise pairs), this seems quite unlikely. Given this, future studies should adopt these changes to the training criteria to avoid spurious training termination and over-exclusion of subjects.

Validation of use of within-subject design: We do not believe that repeated exposure to the TI test could prevent improvement over time on this task. Thus, we

tracked TI ability within-subjects in the present study. This appears to have been validated in Experiment 1. Had there been a learning effect from performing the Inference Test at 20 minutes, we would expect to find some difference when comparing the performance at 3 hours of subjects who had done the Inference Test previously versus those who did not, given how closely the groups were matched on their training parameters; but there was no meaningful difference. Not that this completely rules out whether repeated exposure to the probe trials at the various time points interfered with the normal development in TI ability that might otherwise occur. It is conceivable that once subjects made choices on how to respond to probe trials in the first testing session, they perseverated in responding the same way at later sessions; in effect, learning what their own responses would be at the first testing session, and relying on that memory at later sessions. In a protocol in which subjects would not improve anyway, either because of too little off-line time or because the sleep opportunity was too short, the normal course of lack of improvement would be indistinguishable from this sort of perseveration, and we would see the same results as in Experiment 1. However, this possibility seems far-fetched. Future work could rule this out as a reason for lack of improvement, if they do not find improvement similar to those found in earlier studies (Ellenbogen et al., 2007; Werchan & Gomez, 2013), by looking at whether an early exposure to the Inference Test makes subjects' scores less malleable to later training that is designed to alter responses.

Lack of improvement over the course of a nap: Although we did not see changes in TI performance over the course of 2.5 to 3 hours in Experiment 1, we hypothesized that adding a nap into that period of time could result in changes. Other tasks, in spite of showing off-line improvement over both wake and sleep in some circumstances, show

improvement only over sleep in other circumstances; for example, performance improvements on the MST are only apparent after sleep in non-musicians (Tucker et al., 2016), and improvements on the SRTT are only apparent after sleep when subjects are aware of the sequence (Robertson et al., 2004). Although there are off-line processes that transform memories during wake, a function of sleep may be to make off-line processes more efficient by removing the need to simultaneously remain vigilant of the environment. By shortening the interval between training and test, we may have pushed the opportunity for off-line processes to occur below the threshold at which their effects would become apparent. We could have seen similar improvement on 1º and 2º pairs over a nap, if the difference between sleep and wake is a quantitative difference in the amount of opportunity for off-line processes, and thus a short amount of time were equivalent to a longer time awake; or we could have seen improvement preferentially for the 2º pairs, if sleep and wake support qualitatively difference off-line processes. However, *on average,* performance of subjects in our Nap group did not change between 20 minutes and 3 hours on either 1º or 2º pairs.

This is not because subjects failed to fall asleep. All non-ceiling Nap group subjects got at least 39 minutes of sleep, and at least 18 minutes of N2 sleep. This stage of sleep is characterized by the appearance of sleep spindles, which have been linked to improvement on many tasks. (Bang et al., 2014; Nishida & Walker, 2007) All but one of these subjects got at least 1 minute of N3 sleep, the deepest stage of sleep, and at least 2 minutes of REM sleep, which we hypothesized might have a special role in the integration of disparate information. This is very little time compared to the amount of time usually spent in these stages every night; given this, it is curious that some studies

find a robust effect of naps on task performance, even naps as short as 6 minutes. (Lahl, Wispel, Willigens, & Pietrowsky, 2008) However, the finding that subjects who do not usually take naps do not show any benefit from a nap on a visual perception task (McDevitt, 2014) suggests that a full night of sleep, or nocturnal sleep, may be required for some forms of memory transformation. The TI task may be one of those cases, as our results are in striking contrast to Ellenbogen *et al.* (2007). When they allowed subjects a night of sleep after learning the premise pairs and before testing on the inference pairs, they found these subjects to be, as a group, 20 percentage points better on 1º pairs, and 39 percentage points better on 2º pairs, than subjects tested only 20 minutes after training. However, the contrast between our over-sleep results and Ellenbogen et al.'s could be because our nap study was hampered in its ability to detect improvement by the fact that subjects were already at a high level of performance at baseline.

Difference between Experiment 1 and Experiment 2 baseline performance: Unlike in Experiment 1, quite a large number of subjects were at ceiling at 20 minutes in Experiment 2. The difference in 20 minute TI scores between experiments 1 and 2 was marginally significant ($t(26.1)= 2.1$, $p=0.048$). TI scores were not significantly different from chance at 20 minutes for Experiment 1 subjects, but were extremely significantly higher than chance in Experiment 2. This was problematic for this study, as we were less likely to see improvements in TI ability over sleep if many of the subjects were as good as they could get already before the nap opportunity. We analyzed those who were not at ceiling at 20 minutes to see if the effect of time that Ellenbogen *et al*. surmised perhaps existed amongst a subset of subjects who did not immediately develop TI. Even if we had found something there, it would have been a less interesting finding than Ellenbogen *et*

al.'s claim that subjects, in general, start near chance and that the mean improvement amongst *all* subjects is significant.

Whether there was a reason for this difference between subjects' performances in the two experiments, or whether (noting that the p-value for the disparity is 0.048) this is just the one out of 20 studies to be struck with a glitch of this size, is mysterious. Another difference in subjects' performance between Experiment 1 and Experiment 2 is that the mean number of training blocks required to meet criteria and pass the immediate test was higher in Experiment 2 than in Experiment 1 (Experiment 1: 8.5 ± 0.7; Experiment 2: 14.7 ± 1.7; $t(47.3)= 3.4$, $p=0.001$). However, the increased number of training blocks in Experiment 2 cannot be the confounding variable that mediates the relationship between experiment and initial TI score: in a multiple regression model predicting initial TI score based on number of training blocks and experiment, the coefficient for experiment is much closer to significance ($p=0.053$) than the coefficient for training block count ($p=0.86$). Likewise, a model predicting TI score based on percentage correct on the immediate test of premise pairs and experiment finds experiment to be a significant factor ($p=0.03$) but not the premise pair score ($p=0.49$). The mean score on the immediate test of premise pairs was not different between Experiment 1 and Experiment 2 (Experiment 1: 93% ± 1; Experiment 2: 94% ± 1; $t(61)= 0.25$, $p=0.80$).

The computerized task used to train subjects on the premise pairs was the same for both experiments (except for the changes made partway through Experiment 1, as noted above, which in practice affected only a small number of subjects). Instructions given to the subjects were the same. Subjects in Experiment 1 were run in late fall to winter, whereas subjects in Experiment 2 were run in the spring and summer. It is surprising to

see such a large seasonal effect on a cognitive task in human subjects, but this may have affected who was available amongst the student population, or factors such as how busy they were, or their stress level. The Experiment 1 and Experiment 2 protocols differed in minor ways, which we would not expect to affect subject performance: for example, Experiment 1 subjects had lunch after doing the task, whereas Experiment 2 subjects had lunch immediately before doing the task. All Experiment 2 subjects had EEG electrodes attached to their heads, a tedious process which often took about an hour, which then restricted their movements in a slightly inconvenient way. This, perhaps, set up different expectations amongst subjects regarding how seriously to take the computerized task. On the one hand, whereas it had to be clear to subjects in Experiment 1 that the main focus of the study was the computerized task, subjects in Experiment 2 might have thought that the acquisition of their brain wave signals was their core contribution to the study, and that the computerized task was incidental. For this reason subjects might have taken the task less seriously in Experiment 2, which could explain slower learning of the premise pairs. Also, subjects in Experiment 2 might have been annoyed by the electrode wires, thus reducing attention and motivation on the computerized task. On the other hand, subjects in Experiment 2 might have felt more invested and committed to the study on account of having to undergo the electrode placement, which perhaps resulted greater attention on the Inference Test and higher scores. All of this is completely speculative.

Although subjects were recruited from the same universities, a factor that could have made a large difference biasing who signed up for the study is the wording of the advertisements used to recruit in each experiment. In Experiment 1 the ads made clear that subjects were free to do whatever they wanted between sessions, and most of them

took advantage of this time to study or do homework. The Experiment 2 ads made clear that they would be expected to either nap or watch television between sessions. Most subjects in Experiment 2 complied with this expectation; very few wanted to do their own schoolwork or reading. Thus, the subjects recruited in Experiment 2 may have been those amongst the student population with less pressing academic pressures. Whether a student is in a more or less demanding academic track is largely self-selected, thus we would expect to see differences between the population of subjects with more homework versus the population with less homework. Thus the motivation for not considering Experiment 1 subjects the "Wake" group to compare with "Nap" subjects in Experiment 2: to do a valid comparison between the wake and nap condition, it was imperative to recruit Wake subjects in parallel with Nap subjects, run them at the same time, subject them to the same conditions, and indeed not randomize them into the Nap or Wake condition until the last possible moment. That subjects recruited at different times using different ads would differ somewhat is not surprising, but that the difference was quite so dramatic is surprising.

Introduction of blocked training: Subjects in Experiment 2 were, on average, learning the premise pairs so slowly that an alarming number of the subjects were not managing to meet criteria and pass the immediate test before the 30^{th} block of training. Again we were in the situation of excluding a high percentage of subjects, in spite of the changes to the exit criteria and repetitions of the immediate test discussed above. In previous work it had been observed that animals (and people) learn the premise pairs much more quickly if, rather than have trials of all the premise pairs mixed together randomly, subjects experience massed training on a single premise pair at a time. Rats

and pigeons are almost always trained using this "blocked order" training paradigm. Ellenbogen *et al.* avoided using "blocked order" training, because prior work had found that human subjects, after experiencing blocked order training, were more likely to be able to explicitly state that there was a hierarchy of stimuli, and to be at ceiling in their performance almost immediately. However, the blocked order training, as it had been done in previous studies, had another feature that was probably more responsible for the development of awareness of a hierarchy: In previous studies, the blocks of massed presentations of premise pairs were done *in hierarchy order.* That is, subjects would first experience many trials of A vs. B, followed by many trials of B vs. C, followed by C vs. D, and so on. (This is the "downward" order. Researchers have also done the reverse, "upward" order, i.e. teaching D vs. E first in a five-item series.) Perhaps it is this *ordering* that results in increased awareness, not the massed training. Learning overlapping premise pairs—premise pairs which feature a stimulus in common—in succession may trigger some subjects to observe and consciously reflect on the structure of the task.

Thus, to make training easier for subjects who need it, we introduced a limited amount of a modified blocked order training: training which featured many repetitions of the same premise pair in succession; but unlike previous studies, when training moved to another premise pair, it was always a non-overlapping premise pair. Rather than experiencing premise pairs in the order AB, BC, CD, DE, EF, subjects would experience them in an order such as BC, DE, AB, EF, CD.

Of the subjects who experienced this massed training on premise pairs, some were "aware" according to their responses on the Awareness Questionnaire and others were

not. Overall their awareness scores, and the rate at which they responded “yes” to the hierarchy question, seems to be a bit higher than other subjects in the same experiment, but too few subjects in the present study experienced the massed training to reliably test this statistically. Scores on 2º pairs at the 20 minutes may have been a bit higher with the blocked-order training, but again, more subjects would have to be run to test this. (See Table 5.) Future researchers could consider trying massed training trials in non-hierarchy as an alternative to the standard “downward” order blocked order training, as it may be easier for subjects to learn and yet not lead to any real increase in “awareness”.

	Subjects who experienced blocked-order training	Subjects who experienced random-order training only
Number of included subjects	4	32
Percent who answered Yes to "Did you think any of the patterns were ALWAYS correct (no matter what the other pattern was)?"	75%	75%
Percent who answered Yes to "Did you think any of the patterns were ALWAYS incorrect (no matter what the other pattern was)?"	25%	56%
Percent who answered Yes to "In the test phase, did you notice any new combinations of patterns taken from those you saw before in the training phase?"	100%	56%
Percent who answered Yes to "In the test phase, did you notice any new patterns that you hadn't seen in the training phase?"	100%	63%
Percent who answered Yes to "Did you think there was a hierarchy among the patterns seen in training? That is, do you think they could be ranked from 'best' to 'worst'?"	100%	82%
Mean awareness score	2.0 ± 0.7	1.2 ± 0.2
Percent at ceiling at session 1 on TI pairs	25%	34%
Mean score at 20 minutes on 1º pairs	68% ± 18	69% ± 5
Mean score at 20 minutes on 2º pairs	88% ± 6	82% ± 5

Table 5: Experiment 2, blocked order trained subjects vs. others. Comparison of awareness and performance characteristics subjects included in Experiment 2 who did, or did not, experience the blocked-order training.

Novel-item Test Results: We hypothesized that improvement on 2º pairs relative to 1º pairs reflects an algorithmic change in how the task is done, from a coordination model towards the use of a preference gradient. To test this hypothesis, we need a direct measure of such a preference gradient, independent of subjects' knowledge of relationships within the hierarchy. We hypothesized that the Novel-item Test, by comparing the rate at which subjects chose different items in the hierarchy over an item which could not be logically linked to the hierarchy, would reflect whether a preference

gradient existed, and how strongly it influenced subjects' choices. Thus we hypothesized that there would be some relationship between a subject's Novel-item slope and their relative performance on 2º vs. 1º pairs. This proved to be true in a categorical way, in a non-parametric test: having any tendency to choose higher-ranked items at higher rates over the novel item (as evidenced by a 4-item NiT slope greater than 0) strongly predicted being able to do at least as well on 2º pairs as on 1º pairs. However the magnitude of the slope did not at all correlate with the magnitude of disparity between 2º score and 1º score. This correlation, amongst subjects who had a positive 4-item NiT slope and did at least as well on 2º pairs as 1º pairs, is not significant—and not even positive ($r=-0.17$, $p=0.36$). When the 4-item NiT slope is positive—given that the 4-item NiT slope linearly correlates with overall TI score—just how steep it is may be more of a function of how consistent or noisy a subject's responses are. Thus, in the Novel-item Test, we have a tool for detecting the appearance of a preference gradient, but not for assessing how strong it is.

That the Novel-item Test works is somewhat surprising, as we were concerned that it could be confounded by a tendency to favor familiar stimuli over novel stimuli, or vice versa. Other researchers have assumed that the values of stimuli start off at 0, and thus a novel stimulus would always lose out when compared to any stimulus encountered during TI training, each of which has acquired some association with reward (except perhaps the lowest-ranked item—although it may have acquired some value indirectly via its association with the next-to-lowest-ranked item, according to the Value Transfer Theory.) Thus it is interesting that subjects did choose the novel item fairly frequently. Overriding the association with reward, in some situations animals favor stimuli which are familiar,

for example as shown in the finding that college students rate people they have seen in class more frequently as more attractive; while in other situations the familiar item is disfavored, as in the novel object place recognition task. (Boyce et al., 2016) In the present study, we were concerned, after looking at the data from the first several subjects, that subjects were favoring the familiar stimuli, thus choosing the novel item too infrequently and obscuring any distinctions between items in the hierarchy. We could have added X>Y trials to the training, so that there was prior exposure to these stimuli. Ideally, one would train subjects on X>Y and Y>Z, then have subjects choose between Y and items from the longer hierarchy, so that the "novel" item would be one with some reward history (but not always rewarded). Out of concern for how long it would take to train subjects on even more premise pairs, rather than adding these pairs to training, we merely added instructions telling subjects, before they saw novel items, that a novel item would be the correct choice about half the time. It is unclear, however, what is the effect of attempting to consciously, deliberately override preferences shaped by reward conditioning. Rather than cleanly shifting the bias away from familiar items in an even way that would reveal preferences amongst hierarchy members, erratic attempts to apply this rule may have added noise to the Novel-item Test results.

Gazes et al. (2012), in their experiment 2, attempted something similar to the Novel-item Test with monkeys trained on a 7-item series. In their experiment, monkeys chose between stimuli in the hierarchy and stimuli which had been partially rewarded in another task. Thus they avoided the difficulty of having stimuli that might have been disfavored due to being entirely novel. However, both the task in which monkeys encountered the "novel" stimuli, and the task in which they chose between hierarchy

members and "novel" stimuli, were somewhat different from the transitive inference premise pair training, which might have created the confound that preferences were being tested in a different context. They found that monkeys' preferences amongst the stimuli, as assessed by their rate of choosing hierarchy stimuli over the "novel" items, did not correlate well with order within the hierarchy, according to the Spearman Rank Order correlation. However, this was arrived at by averaging together the responses of all the monkeys. They did not ask whether some monkeys' responses to the hierarchy stimuli, when paired with novel items, corresponded better to the hierarchy order than others, and whether this had any relationship with differences in performances amongst the individual monkeys. It is possible that the monkeys differed in strategy, as did our human subjects, some of whom had perfect or near-perfect transitive performance in spite of 4-item NiT slopes that were close to zero. Averaging together the preferences of a group of subjects may obscure a clear preference gradient that exists in a subset of the subjects. The monkeys had a gradient of responses that decreased from item A to item E, consistent with a preference gradient up to that point, but oddly high responses to items F and G, the two lowest-ranked items. The authors' conclusion was that a preference gradient did not explain performance on the TI task. However, as the same set of stimuli was used for all monkeys, in the same hierarchy order, the possibility that the monkeys had idiosyncratic responses to certain stimuli, obscuring their preference gradient, cannot be ruled out.

Lazereva & Wasserman (2012), using pigeons trained on a five item series, attempted to directly test the values of stimuli B and D in an entirely different way. They theorized that the higher the value of a stimulus to a pigeon, the longer the pigeon's

reactions to the stimulus would show *resistance to extinction*—a continued positive response to the stimulus in the absence of reward—and, after extinction, the less the pigeon's behavior would show *resistance to reinforcement*—a failure to resume positive reactions to the stimulus once reward was reinstated. Previous work had demonstrated that pigeons trained on a single discrimination pair (one rewarded stimulus, and one unrewarded stimulus, always presented together) would peck longer at the rewarded stimulus than at the unrewarded stimulus in extinction phases, during which there were no rewards; and resume pecking at the rewarded stimulus sooner than at the unrewarded stimulus in a reinforcement phase, in which all responses to stimuli were rewarded. However, comparing the pigeons' responses to stimuli B and D from a series of overlapping discriminations revealed no such contrast. Lazereva and Wasserman then went on to alter the relative values of B and D by training the pigeons on many trials of D vs. E. After experiencing many more trials in which stimulus D was rewarded, the value of B was reduced relative to D according to resistance to extinction and resistance to reinforcement results—however accuracy on B vs. D was unchanged! Thus there is a dissociation between TI performance and the elemental value of the stimuli as measured by resistance trials. However, the preference gradient in use during TI trials may be specific to the context of the format of a TI trial; i.e. when aspects of the situation are changed (such as presenting two vs. one stimuli on the pecking screen) the pigeon may revert to relying on the reward frequencies experienced in similar trials, even if they were further in the past.

Given the muddled state of evidence regarding whether performance on the TI task reflects a gradient of preference amongst the stimuli, it would be interesting, in future

work, to attempt a new version of the Novel-item Test that avoids the confounds in the current study (that of the novel items being entirely unfamiliar and never-rewarded) and in Lazereva & Wasserman's (2012) study (that of trying to measure the preferences in a different task context). This could be done by adding X>Y and Y>Z to the TI training and using item Y in the Novel-item Test. As in turned out, the mean training time in Experiment 2 (in which all subjects were trained up to criteria on all the premise pairs, and would re-enter training if they failed the immediate test) was still under 13 minutes amongst subjects who were able to learn the task. Thus the burden of increased training time should not deter future studies from trying this, because the results we have so far with this initial version of the Novel-item Test are intriguing.

As we hypothesized, having a positive Novel-item (NiT) slope—indicating a tendency to choose higher-ranked items more frequently than lower-ranked items over a novel item—is associated with the symbolic distance effect, higher performance on 2º pairs than 1º pairs. This association was significant in a categorical way: the χ^2 test shows that a positive 4-item NiT slope, regardless of how dramatic or how slight, predicts that a subject will perform at least as well on 2º pairs as 1º pairs. However it is not significant as a linear effect: NiT slopes of larger magnitude do not predict a wider spread between 2º and 1º performance. This makes sense, as a preference gradient supports 1º performance as well. So a clearer, stronger preference gradient produces better performance on 2º pairs—but also 1º pairs, so the spread between them does not grow particularly. Indeed, the NiT slope is strongly correlated with both 1º pairs and 2º pairs, consistent with the preference gradient supporting performance of both. Nor is Figure 4.6 linear in the lower left quadrant. Probably, a negative slope generally does not indicate a

reversed preference gradient, but occurs by chance, based on random responses in the absence of any preference differential between the items.

What we did not accomplish in this study, due to time constraints, was to run enough subjects in Experiment 1 who experienced the Novel-item Test at 20 minutes to establish the validity of using repeated-measures with the Novel-item Test. It would be most interesting to see whether the NiT slope changes with time or sleep. But it is imaginable that prior exposure to the Novel-item Test could affect later performance on either test, for example because it could act as a series of extinction trials, or because prior exposure to the novel items would change subjects' relative preferences to them in comparison to the hierarchy members. Ideally it would be best to run additional PR-matched subjects in the 20min*/3hr and 20min/3hr groups, to compare their performance at the later testing session. The second testing session, rather than at 3 hours, should be conducted after a longer delay—even as long as 24 hours (at which point, according to Ellenbogen *et al*, subjects' mastery of the premise pairs is not significantly lower than at 20 minutes). If the off-line improvement that Ellenbogen *et al.* claim to have discovered is real, the question of whether prior exposure to the Novel-item Test disrupts that off-line process is relevant; but we know from this study that 2.5 to 3 hours is too short for the off-line process to occur anyway.

Most intriguingly, we saw differences in how subjects' performance changed over sleep depending on whether their NiT slope was positive. Subsetting the Nap subjects by a threshold NiT slope, we find that average 2º scores went up in one subset and down in the other. This results in close to zero change on average overall amongst the whole group; thus not explaining Ellenbogen *et al.*'s finding that both their groups that slept

before testing had very high 2º scores overall. Perhaps their subjects were more likely to be those who would've had a positive NiT slope, had that test been done; or, less interestingly, the 12-hour sleep and 24-hour groups may have differed as a group from the 20-minute group, and were just better at the 2º pairs overall. Without having within-subject measures, we can only speculate.

Since we did not do the Novel-item Test at the 20 minute testing session, we do not know if NiT slopes would've changed over the nap. Did the subset of Nap subjects with positive NiT slopes at 3 hours, and generally positive improvements on 2º pairs, develop a preference gradient over the course of the nap, which we could've measured as an increase in NiT slope? Or did these subjects have a preference gradient from the start? Perhaps the effect of the nap was to bring subjects' responses on the TI probes more in line with their existing preference gradient. Future work on this task should incorporate the use of the Novel-item Test at more than one point in time, to clarify this.

Although we saw a distinct effect of taking a nap on the performance of subjects who had positive NiT scores, we did not find any correlations between time in particular sleep stages and this effect. The brain is in very different neuromodulatory states in different sleep stages—for example, having high acetylcholine levels during REM, but low during slow-wave sleep—which is thought to facilitate different types of memory processing. Thus we might expect to see sleep-dependent changes in performance occur or not depending on the type of sleep. It has been suggested that the mere initiation of a stage of sleep may produce a cognitive effect (Lahl et al., 2008), and thus it would be interesting to contrast subjects who did or did not get any time in each particular sleep stage. However, with a 90 minute nap, nearly all subjects in the Nap group got at least a

token amount of sleep in each stage, so these any/none contrasts for individual sleep stages are not possible with the current study's data set.

Awareness: Prior work with the Transitive Inference task in humans has established that subjects who are aware of the hierarchical relationship amongst the stimuli perform much better—often at ceiling—on transitive pairs (Smith & Squire, 2005). It has been reported that the "aware" subjects are more likely to rely on a "logic-based" strategy, whereas other subjects rely on "stimulus-driven" strategies; and that the "logic-based" strategy is dependent on declarative memory, which implies awareness (Libben & Titone, 2008). Solomon et al. (2015) put forth the view that TI performance can be supported by conjunctive encoding in the hippocampus, permitting flexible re-combination of the premise pairs (presumably this is the "logic-based" strategy), or by "associative strength-based reinforcement histories of stimuli", mediated by the striatum (the reward system), and that these strategies are competitive. Ellenbogen et al. (2007) reported that improvement on TI over time was not accompanied by an increase in subjects' confidence in their responses; they interpreted this to mean that whatever process occurs over time, it does not involve increasing use of a logic-based strategy, which subjects would presumably be aware of (and thus lead to confidence). Based on these findings, Werchan & Gomez (2013) hypothesized that the off-line, delayed development of TI ability that they were interested in was not supported by the logic-based strategy associated with awareness, and thus they excluded subjects who displayed high awareness in a debriefing.

Based on this, we hypothesized that we would find two different patterns of results if we divided subjects into two subsets, depending on the degree of "awareness" they

displayed on the Awareness Questionnaire: highly aware subjects would be near ceiling on transitive pairs, give highly accurate responses to the questions on the Awareness Questionnaire, and not improve over time; and non-aware subjects would not be at ceiling, but might show changes in performance over time due to an off-line process acting on their preference gradient. As Ellenbogen *et al.*'s finding of no changes in confidence over time seemed to indicate that off-line delayed processes were not leading to insight, we assumed that the answers on the Awareness Questionnaire would reflect the state of subjects' awareness throughout the study, even though it was administered only at the end (and subjects cannot be relied on to accurately remember their prior state of knowledge).

Our findings up-end that picture to some extent. As expected, subjects who answered "yes" to the question "Did you think there was a hierarchy among the patterns seen in training? That is, do you think they could be ranked from 'best' to 'worst'?" performed significantly better on transitive pairs than subjects who answered "no". However, results were inconsistent with the view that these subjects were using a logic-based strategy, and that an "associative strength-based" strategy was competitive to this strategy. Subjects who answered "yes", overwhelmingly, also had stronger preference gradients, as shown on the Novel-item Test. Furthermore, subjects who answered "yes" (and who, before being prompted with the word "hierarchy", gave free-text answers that seemed consistent with awareness of the hierarchy) frequently failed to correctly answer yes-no questions about the task, suggesting insufficient declarative memory about the task to support an explicit logic-based strategy. Thus, an answer of "yes" seemed to

indicate, not that they were using a logic-based strategy, but that *they were using a preference gradient strategy that they were aware of*.

Ideally we should have asked a follow-up question to the "hierarchy" question: of subjects who answered "yes", we should have asked "Must a higher-ranked item ALWAYS win if paired with a lower-ranked item?" As researchers on the Transitive Inference task, we tend to think of the relationship A>B as one with mathematical certainty: if A is taller than B and B is taller than C, then A must, inevitably, be taller than C. However the strategy that subjects use when not told what the relationship is could be one that supports other types of rankings. For example, given three baseball teams, if the Yankees are better than the Twins and the Twins are better than the Cubs, must the Yankees always beat the Cubs? Clearly, no. Hopefully these results lead to a re-thinking of what role "awareness" plays in assessing whether human subjects are using a logic-based strategy.

Exploring the time course of performance on this task can shed light on transitive inference abilities in more ecologically relevant circumstances. Transitive inference experiments have in the past been designed to maximize learning by repeating training trials until a very high level of mastery is displayed, and organizing training trials into repetitive blocks. However, animals evolved to cope with environments which were not designed to optimize learning. For any animal, the greatest survival value would be to make the maximal use of whatever minimal clues the environment provides. This may require using off-line periods and sleep to explore and re-combine waking experiences. An understanding of how the brain's representation of information changes over time could contribute to our understanding of why animals need to sleep.

References

Acuna, B. D., Eliassen, J. C., Donoghue, J. P., & Sanes, J. N. (2002). Frontal and parietal lobe activation during transitive inference in humans. Cereb Cortex, 12(12), 1312-1321.

Acuna, B. D., Sanes, J. N., & Donoghue, J. P. (2002). Cognitive mechanisms of transitive inference. Exp Brain Res, 146(1), 1-10.

Bang, J. W., Khalilzadeh, O., Hamalainen, M., Watanabe, T., & Sasaki, Y. (2014). Location specific sleep spindle activity in the early visual areas and perceptual learning. Vision Res, 99, 162-171.

Barsky, M. M., Tucker, M. A., & Stickgold, R. (2015). REM sleep enhancement of probabilistic classification learning is sensitive to subsequent interference. Neurobiol Learn Mem, 122, 63-68.

Benard, J., & Giurfa, M. (2004). A test of transitive inferences in free-flying honeybees: unsuccessful performance due to memory constraints. Learn Mem, 11(3), 328-336.

Boyce, R., Glasgow, S. D., Williams, S., & Adamantidis, A. (2016). Causal evidence for the role of REM sleep theta rhythm in contextual memory consolidation. Science, 352(6287), 812-816.

Brunamonti, E., Mione, V., Di Bello, F., Pani, P., Genovesio, A., & Ferraina, S. (2016). Neuronal Modulation in the Prefrontal Cortex in a Transitive Inference Task: Evidence of Neuronal Correlates of Mental Schema Management. J Neurosci, 36(4), 1223-1236.

Bryant, P. E., & Trabasso, T. (1971). Transitive Inferences and Memory in Young Children. Nature, 232(52311), 456-458.

Cai, D. J., Mednick, S. A., Harrison, E. M., Kanady, J. C., & Mednick, S. C. (2009). REM, not incubation, improves creativity by priming associative networks. Proc Natl Acad Sci U S A, 106(25), 10130-10134.

Conte, F., & Ficca, G. (2013). Caveats on psychological models of sleep and memory: a compass in an overgrown scenario. Sleep Med Rev, 17(2), 105-121.

Daniels, C. W., Laude, J. R., & Zentall, T. R. (2014). Six-term transitive inference with pigeons: successive-pair training followed by mixed-pair training. J Exp Anal Behav, 101(1), 26-37.

Delgado, M. R., & Dickerson, K. C. (2012). Reward-related learning via multiple memory systems. Biol Psychiatry, 72(2), 134-141.

Delius, J. D., & Siemann, M. (1998). Transitive responding in animals and humans: Exaptation rather than adaptation? Behavioural Processes, 42, 107-137.

DeVito, L. M., Lykken, C., Kanter, B. R., & Eichenbaum, H. (2010). Prefrontal cortex: role in acquisition of overlapping associations and transitive inference. Learn Mem, 17(3), 161-167.

Dusek, J. A., & Eichenbaum, H. (1997). The hippocampus and memory for orderly stimulus relations. Proc Natl Acad Sci U S A, 94(13), 7109-7114.

Eichenbaum, H., & Fortin, N. J. (2009). The neurobiology of memory based predictions. Philos Trans R Soc Lond B Biol Sci, 364(1521), 1183-1191.

Ellenbogen, J. M., Hu, P. T., Payne, J. D., Titone, D., & Walker, M. P. (2007). Human relational memory requires time and sleep. Proc Natl Acad Sci U S A, 104(18), 7723-7728.

Frank, M. J., Rudy, J. W., Levy, W. B., & O'Reilly, R. C. (2005). When logic fails: implicit transitive inference in humans. Mem Cognit, 33(4), 742-750.

Frank, M. J., Rudy, J. W., & O'Reilly, R. C. (2003). Transitivity, flexibility, conjunctive representations, and the hippocampus. II. A computational analysis. Hippocampus, 13(3), 341-354.

Gallagher, M. (1990). Introduction. In J. L. McGaugh, N. M. Weinberger, & G. Lynch (Eds.), Brain Organization and Memory: Cells, Systems, and Circuits. New York: Oxford University Press.

Gazes, R. P., Chee, N. W., & Hampton, R. R. (2012). Cognitive mechanisms for transitive inference performance in rhesus monkeys: measuring the influence of associative strength and inferred order. J Exp Psychol Anim Behav Process, 38(4), 331-345.

Gazes, R. P., Lazareva, O. F., Bergene, C. N., & Hampton, R. R. (2014). Effects of spatial training on transitive inference performance in humans and rhesus monkeys. J Exp Psychol Anim Learn Cogn, 40(4), 477-489.

Greene, A. J. (2007). Human hippocampal-dependent tasks: is awareness necessary or sufficient? Hippocampus, 17(6), 429-433.

Greene, A. J., Gross, W. L., Elsinger, C. L., & Rao, S. M. (2006). An FMRI analysis of the human hippocampus: inference, context, and task awareness. J Cogn Neurosci, 18(7), 1156-1173.

Greene, A. J., Spellman, B. A., Dusek, J. A., Eichenbaum, H. B., & Levy, W. B. (2001). Relational learning with and without awareness: transitive inference using nonverbal stimuli in humans. Mem Cognit, 29(6), 893-902.

Gupta, A. S., van der Meer, M. A., Touretzky, D. S., & Redish, A. D. (2010). Hippocampal replay is not a simple function of experience. Neuron, 65(5), 695-705.

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)--a metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform, 42(2), 377-381.

Higa, J. J., & Staddon, J. E. R. (1993). "Transitive Inference" in Multiple Conditional Discriminations. Journal of the Experminental Analysis of Behavior, 59(2), 265-291.

Hoddes, E., Zarcone, V., Smythe, H., Phillips, R., & Dement, W. C. (1973). Quantification of sleepiness: a new approach. Psychophysiology, 10(4), 431-436.

Iber, C., Ancoli-Israel, S., Chesson, A., & Quan, S. (2007). The AASM Manual for the Scoring of Sleep and Associated Events: Rules, Terminology and Technical Specifications (1st ed.). Westchester, Illinois: American Academy of Sleep Medicine.

Jacobs, L. F. (2006). From movement to transitivity: the role of hippocampal parallel maps in configural learning. Rev Neurosci, 17(1-2), 99-109.

Jenkins, J., & Dallenbach, K. (1924). Obliviscence During Sleep and Waking. American Journal of Psychology, 35, 605–612.

Johns, M. W. (1991). A new method for measuring daytime sleepiness: the Epworth sleepiness scale. Sleep, 14(6), 540-545.

Kumaran, D., & Ludwig, H. (2013). Transitivity performance, relational hierarchy knowledge and awareness: results of an instructional framing manipulation. Hippocampus, 23(12), 1259-1268.

Kumaran, D., & McClelland, J. L. (2012). Generalization through the recurrent interaction of episodic memories: a model of the hippocampal system. Psychol Rev, 119(3), 573-616.

Lahl, O., Wispel, C., Willigens, B., & Pietrowsky, R. (2008). An ultra short episode of sleep is sufficient to promote declarative memory performance. J Sleep Res, 17(1), 3-10.

Lakoff, G., & Johnson, M. (1980). Metaphors We Live By. Chicago: University of Chicago Press.

Lazareva, O. F., Smirnova, A. A., Bagozkaja, M. S., Zorina, Z. A., Rayevsky, V. V., & Wasserman, E. A. (2004). Transitive responding in hooded crows requires linearly ordered stimuli. J Exp Anal Behav, 82(1), 1-19.

Lazareva, O. F., & Wasserman, E. A. (2006). Effect of stimulus orderability and reinforcement history on transitive responding in pigeons. Behav Processes, 72(2), 161-172.

Lazareva, O. F., & Wasserman, E. A. (2010). Nonverbal transitive inference: Effects of task and awareness on human performance. Behav Processes, 83(1), 99-112.

Lazareva, O. F., & Wasserman, E. A. (2012). Transitive inference in pigeons: measuring the associative values of Stimuli B and D. Behav Processes, 89(3), 244-255.

Leo, P. D., & Greene, A. J. (2008). Is awareness necessary for true inference? Mem Cognit, 36(6), 1079-1086.

Libben, M., & Titone, D. (2008). The role of awareness and working memory in human transitive inference. Behav Processes, 77(1), 43-54.

Mackey, A. P., Miller Singley, A. T., Wendelken, C., & Bunge, S. A. (2015). Characterizing Behavioral and Brain Changes Associated with Practicing Reasoning Skills. PLoS One, 10(9), e0137627.

Marshall, L., & Born, J. (2007). The contribution of sleep to hippocampus-dependent memory consolidation. Trends Cogn Sci, 11(10), 442-450.

Martin, N., & Alsop, B. (2004). Transitive inference and awareness in humans. Behav Processes, 67(2), 157-165.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. Psychol Rev, 102(3), 419-457.

McDevitt, E. A., Whitehurst, L., Duggan, K, and Mednick, S. (2014). Individual differences in sleep-dependent perceptual learning: Habitual vs. non-habitual nappers. Journal of Vision, 14(10), 1176.

McKeon, S., Pace-Schott, E. F., & Spencer, R. M. (2012). Interaction of sleep and emotional content on the production of false memories. PLoS One, 7(11), e49353.

Merritt, D. J., & Terrace, H. S. (2011). Mechanisms of inferential order judgments in humans (Homo sapiens) and rhesus monkeys (Macaca mulatta). J Comp Psychol, 125(2), 227-238.

Mukhametov, L., Supin, A., & Polyakova, I. (1977). Interhemispheric asymmetry of the electroencephalographic sleep patterns in dolphins. Brain Research, 134, 581-584.

Nguyen, N. D., Tucker, M. A., Stickgold, R., & Wamsley, E. J. (2013). Overnight Sleep Enhances Hippocampus-Dependent Aspects of Spatial Memory. Sleep, 36(7), 1051-1057.

Nishida, M., & Walker, M. P. (2007). Daytime naps, motor memory consolidation and regionally specific sleep spindles. PLoS One, 2(4), e341.

Payne, J. D., Stickgold, R., Swanberg, K., & Kensinger, E. A. (2008). Sleep preferentially enhances memory for emotional components of scenes. Psychol Sci, 19(8), 781-788.

Robertson, E. M., Pascual-Leone, A., & Press, D. Z. (2004). Awareness modifies the skill-learning benefits of sleep. Curr Biol, 14(3), 208-212.

Schlichting, M. L., & Preston, A. R. (2015). Hippocampal-medial prefrontal circuit supports memory updating during learning and post-encoding rest. Neurobiol Learn Mem.

Sio, U., Monaghan, P., & Ormerod, T. (2012). Sleep on it, but only if it is difficult: Effects of sleep on problem solving. Mem Cognit.

Smith, C., & Squire, L. R. (2005). Declarative memory, awareness, and transitive inference. J Neurosci, 25(44), 10138-10146.

Solomon, M., Ragland, J. D., Niendam, T. A., Lesh, T. A., Beck, J. S., Matter, J. C., . . . Carter, C. S. (2015). Atypical Learning in Autism Spectrum Disorders: A

Functional Magnetic Resonance Imaging Study of Transitive Inference. J Am Acad Child Adolesc Psychiatry, 54(11), 947-955.

Spitzer M., W. J., Rappsilber J., Tyers M. (2014). BoxPlotR: a web tool for generation of box plots. Nature Methods, 11(2), 121-122.

Stickgold, R., Scott, L., Rittenhouse, C., & Hobson, J. A. (1999). Sleep-induced changes in associative memory. J Cogn Neurosci, 11(2), 182-193.

Stickgold, R., Whidbee, D., Schirmer, B., Patel, V., & Hobson, J. A. (2000). Visual discrimination task improvement: A multi-step process occurring during sleep. J Cogn Neurosci, 12(2), 246-254.

Tucker, M. A., Nguyen, N., & Stickgold, R. (2016). Experience Playing a Musical Instrument and Overnight Sleep Enhance Performance on a Sequential Typing Task. PLoS One, 11(7), e0159608.

Van Elzakker, M., O'Reilly, R. C., & Rudy, J. W. (2003). Transitivity, flexibility, conjunctive representations, and the hippocampus. I. An empirical analysis. Hippocampus, 13(3), 334-340.

Vasconcelos, M. (2008). Transitive inference in non-human animals: an empirical and theoretical analysis. Behav Processes, 78(3), 313-334.

von Fersen, L., Wynne, C. D. L., & Delius, J. D. (1991). Transitive Inference Formation in Pigeons. Journal of Experimental Psychology, 17(3), 334-341.

Walker, M. P., Brakefield, T., Morgan, A., Hobson, J. A., & Stickgold, R. (2002). Practice with sleep makes perfect: sleep-dependent motor skill learning. Neuron, 35(1), 205-211.

Walker, M. P., & Stickgold, R. (2010). Overnight alchemy: sleep-dependent memory evolution. Nat Rev Neurosci, 11(3), 218; author reply 218.

Wendelken, C., & Bunge, S. A. (2010). Transitive inference: distinct contributions of rostrolateral prefrontal cortex and the hippocampus. J Cogn Neurosci, 22(5), 837-847.

Werchan, D. M., & Gomez, R. L. (2013). Generalizing memories over time: sleep and reinforcement facilitate transitive inference. Neurobiol Learn Mem, 100, 70-76.

Xie, L., Kang, H., Xu, Q., Chen, M. J., Liao, Y., Thiyagarajan, M., . . . Nedergaard, M. (2013). Sleep drives metabolite clearance from the adult brain. Science, 342(6156), 373-377.

Zeithamova, D., Schlichting, M. L., & Preston, A. R. (2012). The hippocampus and inferential reasoning: building memories to navigate future decisions. Front Hum Neurosci, 6, 70.